Charles University

Faculty of Science

Department of Physical and Macromolecular Chemistry

Study program: Biophysical chemistry

Study of structural features of single stranded DNA by biophysical techniques and crystallography

Studie strukturních vlastností jednovláknových DNA biofyzikálními metodami a krystalograficky

Diploma thesis/Diplomová práce

Supervisor: Prof. Ing. Bohdan Schneider, CSc., DSc

Praha 2021 Bc. Jakub Svoboda

Prohlášení

Prohlašuji, že jsem tuto diplomovou práci vypracoval samostatně pod vedením svého školitele prof. Ing. Bohdana Schneidera, CSc., DSc a všechny použité prameny jsem řádně citoval. Práce nebyla využita jako závěrečná práce k získání jiného nebo obdobného druhu vysokoškolské kvalifikace.

V Praze, dne ……

……..

Bc. Jakub Svoboda

Poděkování

Předně bych chtěl poděkovat svému školiteli prof. Ing. Bohdanu Schneiderovi CSc., DSc za vědeckou výchovu a možnost vypracovat tuto práci v přátelském pracovním prostředí. Dále bych chtěl poděkovat Tatsiane Charnavets Ph.D. za naměření spekter cirkulárního dichroismu a doc. Ing. Petru Kolenkovi Ph.D. za naměření difrakčních dat a vedení při řešení krystalových struktur. V neposlední řadě bych chtěl poděkovat všem blízkým za podporu během studia.

Práce vznikla pod záštitou následujících grantů:

LTAUSA18197 Design, development, and testing of bioinformatic tools for validation of experimental and computer molecular models in structural biology, biotechnology and pharmacy.

CIISB4HEALTH This publication was supported by the project Czech infrastructure for integrative structural biology for human health (CZ.02.1.01/0.0/0.0/16_013/0001776) from the ERDF.

Tuto práci věnuji své babičce Mirce

Abstract

DNA is the fundamental molecule in all domains of life, its role in heredity is well established. Although the famous double helical complementary form is indispensable for replication mechanism DNA can occupy wide range of conformations. In the past studies performed in the laboratory, DNA oligomers related to single stranded bacterial Repetitive Extragenic Palindromic (REP) showed spectral behavior suggesting complex equilibria including double helical, hairpin, and tetraplex conformations. The studies presented in this thesis extended the scope of analyzed sequences and employed circular dichroism spectroscopy and X-ray crystallography. We report spectral data and X-ray structures of three successfully crystalized oligonucleotides. All three structures acquire double helical architecture with two consecutive T-T mismatches in the center. To improve the convergence of the refinement process of the crystal structures we used novel dinucleotide conformational classes, NtC classes. The NtC class classification was also used to analyze geometries of selected non-canonical base pairs in all DNA crystal structures in the Protein Data Bank. We measured the fit between geometries of the dinucleotides involved in the non-canonical base pairing and the NtC classes and correlated this fit to the electron density of the analyzed dinucleotides. This new type of the quality measure revealed that dinucleotides involved in non-canonical base pairing show no significant geometrical difference from the Watson-Crick paired dinucleotides. We also suggest that a large fraction of so far unclassified dinucleotides can be re-refined into the known geometries.

Key words: single stranded DNA, structural database, X-ray crystallography, circular dichroism, NtC, base pairing, mismatch

Abstrakt

Ve všech doménách života je DNA základní molekolou, její úloha v dědičnosti je dobře etablovaná. Ačkoliv její proslavená dvojšroubovicová komplementární forma je nenahraditelná pro replikační mechanismus, může i tak zaujímat širokou škálu konformačních rodin. V dřívějších pracích studované jednovláknové bakteriálním Repetitivním Extragenním Palindromickým (REP) sekvencím příbuzné DNA oligomery vykazovali komplexní spektrální profil zahrnující dvojšroubovice, vlásenky a tetraplexy. Studie předkládané v této práci rozšiřují měřítko analyzovaných sekvencí a mapují konformační prostor s REP příbuznými oligonukloetidy s použitím cirkulárního dichroismu a krystalografie. Spektrální data a krystalové struktury tří oligonukleotidů jsou reportovány. Všechny tři varianty krystalizovali do duplexové formy se dvěma po sobě jdoucími T-T páry v centrální části. Pro vylepšení rafinačního procesu krystalových struktur byly použity nové dinukleotidové konformační třídy, NtC. Klasifikace pomocí NtC byla použita také k analýze vybraných nekanonických párů v krystalových strukturách získaných z PDB. Následně byl měřen fit mezi geometrií nukleotidů zapojených v nekanonických párech a NtC třídami, tento fit byl dále korelován s elektronovou hustotou analyzovaného dinukleotidu. Tento nový typ měřítka kvality odhalil, že dinukleotidy zapojené v nekanonickém párování nevykazují signifikantní geometrický rozdíl od dinukleotidů párovaných Watson-Crickovským typem. Dále navrhujeme, že velká část zatím nepřiřazených dinukleotidů může být znovu rafinována do známe NtC třídy.

Klíčová slova: jednovláknová DNA, strukturní databáze, krystalografie, cirkulární dichroismus, NtC, párování bazí, mismatch

Table of contents

Abbreviations

BrU - Bromouracil

CANA – Conformational alphabet of nucleic acids

CD – Circular dichroism

mmCIF - Macromolecular crystallographic information file

MPD - 2-Methylpentane-2,4-diol

NtC – diNucleotide conformer

PDB – Protein data bank

RAYT – REP associated tyrosine transposase

REP – Repetitive extragenic palindromes

r.m.s.d. – Root mean square deviation

RSCC - Real-space correlation coefficient

SVD – Singular value decomposition

Introduction

Nucleic acids are biologically significant polymers involved in the preservation and expression of the genetic information. This flow of genetic information was termed the central dogma of molecular biology. While DNA is primarily employed in preserving the genetic information in its sequence which is precisely maintained by molecular machinery, RNA has a multiple functional roles in form of transfer RNA and ribosome complex, both involved in protein synthesis. It has been reported that some nucleic acid molecules, so called ribozymes, show catalytic activity when splicing RNA in gene expression. Discoveries in the field of processive enzymes - polymerases opened a floodgate of molecular techniques based on polymerase chain reaction which found routine uses in basic research and is the base of many diagnostic methods.

The overall dominant helical structure of DNA and sophisticated architecture of RNA have been probed by various techniques. Atomic resolved X-ray structures have helped our understanding of nucleic acids and their relationships to heredity and life. Their dynamic behavior is described by spectroscopic studies such as nuclear magnetic resonance or fluorescence techniques.

Recent development in experimental techniques such as cryo-electron microscopy, optical and magnetic tweezers or atomic force microscopy have enabled rapid expansion of the DNA bionanotechnology field. Apparent readiness for programmability of the strand annealing through its complementarity led to development of a variety of algorithms which are used to predict folding patterns of large structures in the DNA origami-like manner.

History

Isolation and characterization of nucleic acids

Discovery and characterization of nucleic acids represent several milestones in the field of life sciences. Study of nucleic acids began in 1869 with their isolation from white blood cells. The discovery was made by Swiss scientist Friedrich Miescher who called the isolated substance a nuclein. Due to its acidic properties the name was later changed to nucleic acid (Nelson 2005, Neidle 2008). After decades of hardly any progress Oswald Avery, Colin MacLeod and Maclyn McCarthy published a paper in 1944 in which they demonstrated that genes were made of nucleic acids (Avery, MacLeod et al. 1944). Their results were further confirmed by Alfred Hershey and Martha Chase in 1952 (Hershey and Chase 1952). Investigation of DNA composition in different origin specimens led Erwin Chargaff to discover that molar ratios of cytosine are equal to guanine and that of adenine to thymine. This has later become basis for Chargaff rules (Zamenhof, Brawermann et al. 1952).

The fiber diffraction experiment

Famous photo 51 (Figure 1), taken by Raymond Gosling, graduate student working under the supervision of Rosalind Franklin and Maurice Wilkins, shows X-ray diffraction of gel composed of DNA fibers. The diffraction pattern is believed to be crucial for developing the atomistic model of double helical structure of DNA. Franklin and Gosling commented that the relative humidity has a significant role in overall structure, images with different relative humidities were named as structure A and B, later leading to two forms of nucleic acids termed A-form and B-form respectively (Franklin and Gosling 1953).

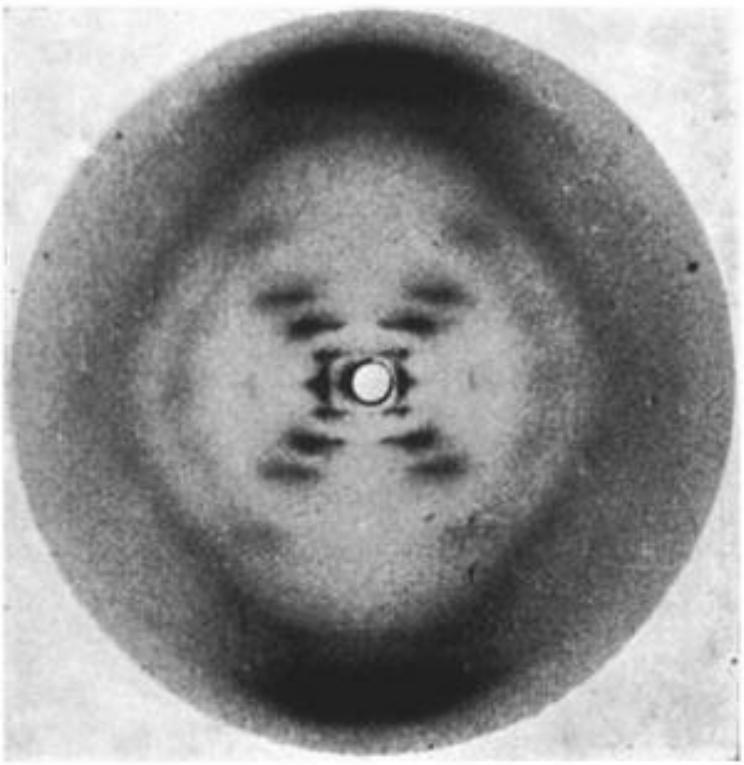

Figure 1: Photo 51 taken by Raymond Gosling under supervision of Rosalind Franklin at King's College London in 1952 has become influential source of information for successful building of the atomic model of DNA double helix (Franklin and Gosling 1953).

In 1953, Watson and Crick pieced together all available information, mainly Chargaff rules and X-ray diffraction images, and came up with the now famous wire model of DNA double helix (Watson and Crick 1953). For building the correct model and the realization that strands are complementary and this complementarity is the base for replication the trio Watson, Crick and Wilkins were awarded with a Nobel Prize in physiology and medicine (Crick and Watson 1954).

Structure of nucleic acids

Nucleotide

Phosphate group, aromatic base and (deoxy)ribose are main elements of nucleotides, fundamental building blocks of nucleic acids. Nucleic acids naturally occur in the form of ribonucleic (RNA) or deoxyribonucleic acid (DNA), depending on whether the ribose or deoxyribose is present. Phosphate groups and (deoxy)ribose are linked by phosphodiester bonds creating linear sugar-phosphate backbone in the process. Bases are attached to sugar rings via glycosidic bonds formed between C1´ of sugar and N1 of pyrimidine or N9 of purine base (Nelson 2005, Neidle 2008).

The bases are planar aromatic heterocyclic compounds and are divided into two groups: the single-ring pyrimidine bases cytosine, thymine and uracil and the double-ring purine bases adenine and guanine (Figure 2). It is widely used practice to write nucleic acid sequences in single-letter code - C, T/U, A, G. Bases that are derived from pyrimidine rings, C and T/U can be abbreviated with letter Y, purine bases, G and A with letter R (Nelson 2005, Neidle 2008).

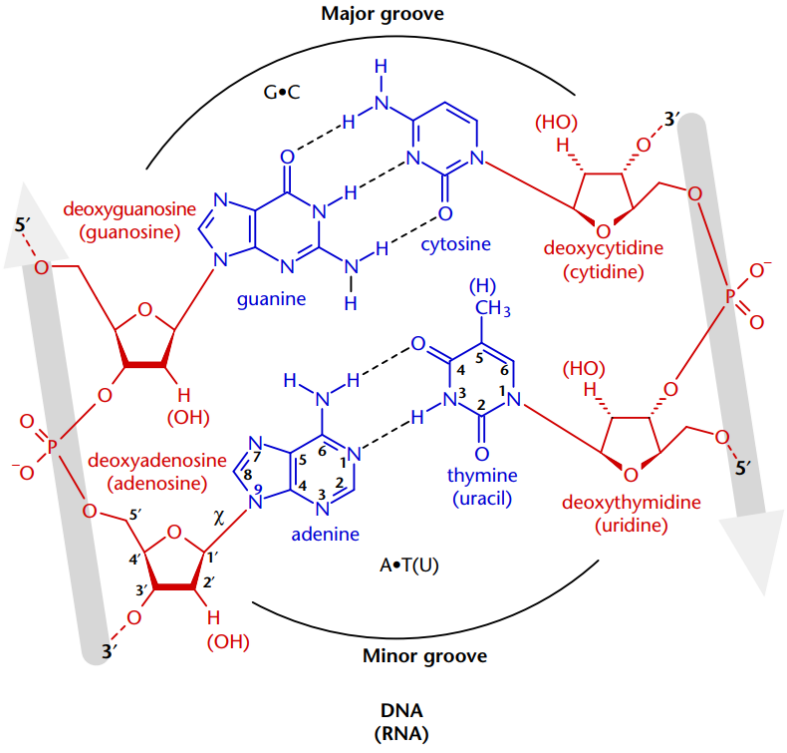

Figure 2: Major bases of nucleic acids and their hydrogen bonding pattern, directionality of chains is indicated as well (Soukup 2003).

They can be found in two tautomeric forms, both coexist in the equilibrium with each other. Nitrogen atoms are preferentially in the amino form rather than imino, same applies for oxygen atoms and their enol and keto configuration, in which keto form is preferred. Minor enol and imino forms are responsible for a significant number of errors during replication of the DNA (Neidle 2008, Singh, Fedeles et al. 2015).

Due to an aromatic character they are able to absorb electromagnetic waves with λ_{max} values around 260 nm and generate a characteristic spectrum which is important in the experimental detection of nucleic acids. When rings of the bases are structured in a face to face orientation, e.g. in the polynucleotide chain, UV absorption is reduced due to a sharing of π-electrons, which profoundly changes the transition dipoles of the bases. This effect is known as hypochromicity (Cox 1970). The hypochromicity is often used in temperature dependent experiments such as calorimetric measurements since it gives information on changes in base stacking. While being hydrophobic, they readily form hydrogen bonds via utilizing substituents on the ring's edges, which is crucial for binding small molecules, ligands, proteins and other nucleic acids (Blackburn and Gait 2006, Neidle 2008).

Since the (deoxy)ribose has uniquely described numbering of atoms, it is used to define the direction (polarity) of the nucleic acid strand as 5′ to 3′ end of each strand. Sp^3 hybridization on the carbon atoms causes nonplanarity of the entire sugar ring. This intrinsic property is termed puckering. The conformation of the ring can be described with five torsion angles τ_0-τ_4, which are dependent on each other. There are, in principle, many puckers separated by energy barriers. Indeed, multitude of distinct deoxyribose puckers have been observed by NMR and X-ray diffraction techniques. The puckers can be defined using parameters P a τ_m, P is the phase angle of pseudorotation and it indicates the type of pucker. It can take any value between 0° and 360°. τ_m is the maximum degree of the pucker, experimental values obtained from the crystal structures of mononucleosides are in the range between 25° and 45° (Neidle 2008).

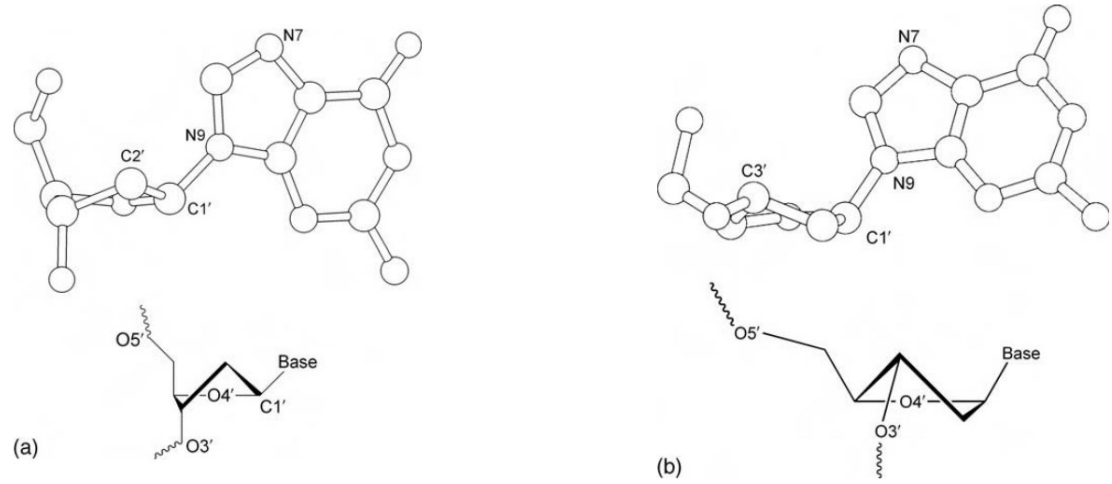

Figure 3: Two most important sugar pucker modes. (a) Shows C2'-endo sugar pucker and (b) C3'-endo conformation (Neidle 2008).

Two of the most common puckers are C2′-endo and C3′-endo, they differ in whether the C2′ or C3′ are on the same side as the base and C4′-C5′ bond (Figure 3). If one sugar atom is under the plane defined by other sugar atoms it is called exo. Usually, when one atom is relatively more deviated from the plane to one side, another atom is slightly deviated to the other side. C2′-endo puckers have values of P in the range of 140-185°, C3′-endo in the range of -10 to +40°. Different populations of puckers can be observed in NMR experiments, these show that in solution the interconversion of puckers occurs swiftly. Population of the major pucker is a base type dependent. C2′-endo conformation correlates with the purines and C3′-endo with pyrimidines. In order to better visualize the phenomenon of sugar pucker, it is valuable to represent pseudorotation phase angle in the form of a conformational wheel shown in Figure 4 (Blackburn and Gait 2006, Neidle 2008).

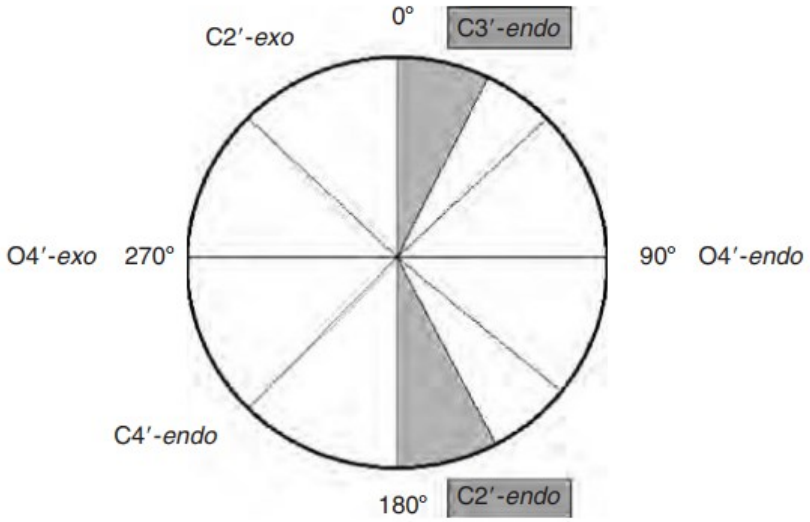

Figure 4: The pseudorotation wheel of a deoxyribose sugar. Preferred geometries are indicated by the shaded area (Neidle 2008).

The glycosidic bond connects deoxyribose sugar and a base. Torsion angle of this bond, χ, is defined with atoms O4′-C1′-N9-C4 for purines and O4′-C1′-N1-C2 for pyrimidines. There are two theoretically predicted low-energy positions for glycosidic angle - anti and syn. The anti-conformation has the N1, C2 face of purines and C2, N3 face of pyrimidines positioned away from the sugar ring. The syn conformation has this orientation reversed. Nucleotides containing guanine accepts the syn orientation (Blackburn and Gait 2006, Neidle 2008).

Sugar-phosphate backbone has six variable torsion angles - α, β, γ, δ, ε and ζ in addition to the five sugar torsion angles τ_0-τ_4 and glycosidic angle χ (backbone and glycosidic torsions are shown in Figure 5). Many of these angles have highly correlated values. Since each of the torsion angles has some degree of the steric freedom, consequently there are many low-energy conformers for the nucleotide unit (Schneider, Boaeikova et al. 2018, Cerny, Bozikova et al. 2020).

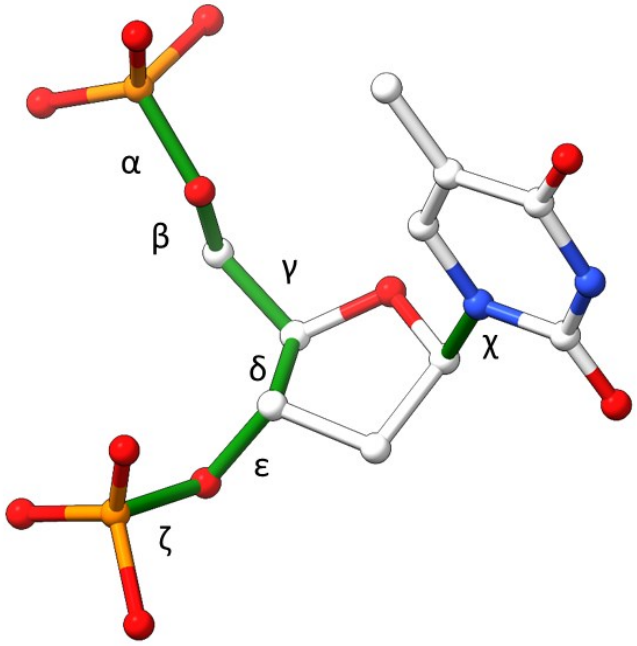

Figure 5: The nucleotide backbone torsion angles (α, β, γ, δ, ε and ζ) and glycosidic torsion χ are depicted (Nelson 2005). Figure was created in ChimeraX (version 0.92) (Pettersen, Goddard et al. 2021).

Hydration of nucleic acids

Together with base, pentose and phosphate, hydration is often called the fourth fundamental part of nucleic acid structure to emphasize the impact of organized water molecules in close proximity to the DNA/RNA molecule (Westhof 1988). Water molecules are not just a solvent but play a significant role in dynamics and interactions with other molecules (for example - catalysis of electron transfer reaction) (Blackburn and Gait 2006).

Water helps nucleic acids to fold and stabilize three-dimensional structure. All-atom molecular simulations revealed that folding occurs through hydrophobic collapse and expulsion of solvent from the core part. Water also mediates minor groove interaction with other binders via creating interface thus affecting overall dynamics of the interaction process (Neidle 2008).

Crystallographic studies give the most precise experimental insight into hydration of biomolecules. While being extremely difficult, for detailed analysis crystals of high resolution (around or below 1 Å) must be obtained. It is beneficial to see hydration in crystal structures as a time-averaged hydration shell rather than one static image (Schneider and Berman 1995, Schneider, Patel et al. 1998). Hydration of nucleic acids is type and sequence dependent. This fact can be used in prediction of binding sites (Biedermannova and Schneider 2016).

Step Base-Base interactions

While the base pairing, mainly its complementarity, is the key to the copying mechanism that ensures the integrity of the genetic information of the host organism, it is only a part of the overall energy landscape (Parker, Hohenstein et al. 2013). Step base-base interaction, the interaction driving the DNA and RNA folding, along with hydrophobic effect play a pivotal role in this regard (Neidle 2008, Fallmann, Will et al. 2017).

There are two types of the step interaction that have an effect on the geometry of the helix. First is the repulsive interaction between bases caused by steric interactions between methyl groups on thymine, the guanine amino groups and the configuration of the step. Second interaction is a stacking interaction that consists of a van der Waals component and series of electrostatic interactions between partial charges and between the charge distributions associated with the π electron density above and below bases (Blackburn and Gait 2006).

Base Pairing

The nitrogen and oxygen containing substituents on the base rings and the heteroatoms inside of the rings themselves are either donor or acceptor of hydrogen bonds. In this section hydrogen bonds formed between two or multiple bases will be discussed in greater detail. As mentioned above, the particular geometry of hydrogen donors and acceptors on the edges of bases is the main driving force for forming hydrogen bonds. NH groups are good hydrogen bond donors and oxygen on the carbonyl group and free electron pair on ring nitrogens are both hydrogen bond acceptors. Pairing proposed by Watson between cytosine and guanine can be viewed as a pattern of a.a.d on cytosine and d.d.a on guanine (Figure 6), similar scheme of donors and acceptors of hydrogen bonds can be constructed on adenine and thymine base pair. It has been shown that specificity of interaction between nucleic acids and other molecules is achieved through directionality of hydrogen bonding pattern on bases (Seeman, Rosenberg et al. 1976, Blackburn and Gait 2006).

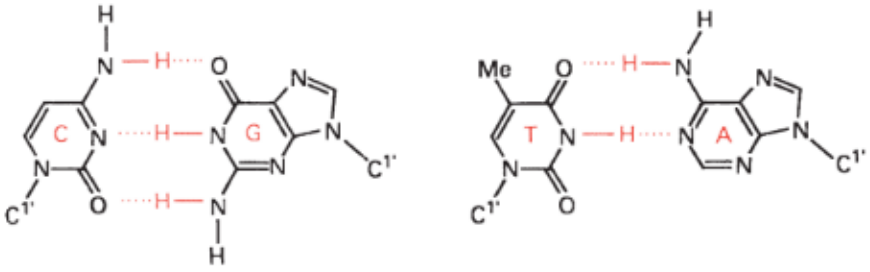

Figure 6: Hydrogen donor and acceptor pattern in canonical Watson-Crick base pair (Seeman, Rosenberg et al. 1976, Blackburn and Gait 2006).

Morphology of the individual base pairs is to a great extent flexible. This flexibility is dependent on the nature of bases and their environment, more specifically conditions surrounding the

molecule or stacking interaction between bases under and above the examined base pair. A lot of base pair parameters have been defined gradually over the past years, some of them will be discussed (Neidle 2008).

Details of the DNA double helical architecture are traditionally described by various parameters, Figure 7. Base pairs often deviate from planarity, partly because angle between atoms involved in a hydrogen bond can deviate up to 35 degrees and partly to avoid steric clashes with other non-bonded bases. The non-planarity can be defined as movement between two base pairs (e.g. propeller twist, buckle or stretch) or as movement between base pairs relative to the helix axis (displacement in the plane of base pair or inclination). The second way to look at the distortion from ideal values is from point of view of successive base pairs - base pair step. Examples for later are helical twist, roll, rise etc. (Neidle 2008).

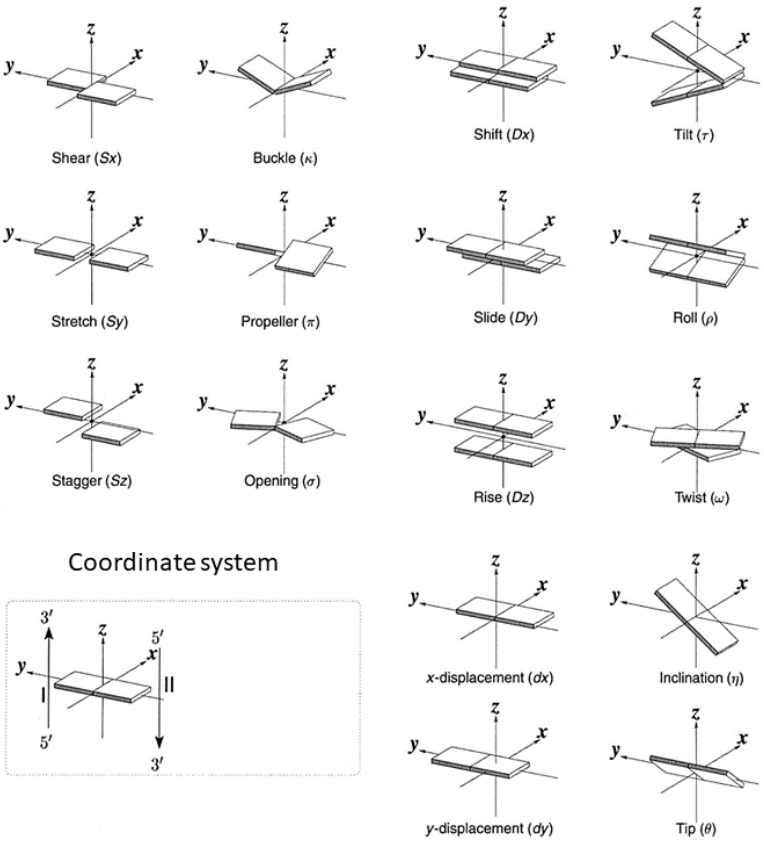

Figure 7: Examples of base pair and base step parameters (Neidle 2008).

As a consequence of the asymmetry of base pairs in the helix, major and minor groove can be distinguished. Conventionally, the bonds C1'-N9 of purines and C1'-N1 of pyrimidines are on the minor groove side. Dimensions describing major and minor groove can vary largely between different structural forms of nucleic acids. Grooves are quite important because many classes of proteins interact with nucleic acids through minor or major groove interface (Neidle 2008).

Canonical and non-canonical base pairing

Two canonical base pairs A-T and C-G mostly occur as the result of isostericity, they display least sterical stress along torsions in sugar phosphate backbone and finally acceptors and donors are in the optimal distance from each other (Schneider and Berman 2006). They obey the complementarity rule, which is a key component for preservation of the genetic information (Nelson 2005, Neidle 2008).

However, it has been observed and also theoretically predicted that other than canonical Watson-Crick hydrogen bond geometries are possible. They play a key role in the building of some three dimensional DNA structures such as multistranded tetraplexes - G-tetraplexes or cytosine rich i-motif, triple helices, or various looped motifs. The non-canonical pairing is more frequent in RNA structures, where it is necessary for allowing more complex architectures to arise. In contrast, non-canonical base pairs are said to have a destabilizing effect on the DNA duplex. Two of the most widely used descriptions using Watson-Crick, Hoogsteen and sugar edge were developed by Leontis/Westhof and Saenger (Saenger 1984, Leontis and Westhof 2001). Nomenclature proposed by duo Leontis and Westhof has 16 combinations, Saenger defined 28 classes (Figure 8). Both methods can be used simultaneously since they use different protocols to classify base pairs. Leontis and Westhof, while being widely used in RNA structures, used a description of edges, while Saenger took a more descriptive atomistic interpretation with hydrogen bond positions (Saenger 1984, Neidle 2008).

Synthetic or modified bases enrich structural variability of already plastic nucleic acids, for example they can lock bases in such conformations that they can bind to other strand via Hoogsteen edge. Replacement of adenine by 2-oxo-adenine reduces electrostatic repulsion in Hoogsteen geometry, while maintaining a number of hydrogen bonds and leaving stacking efficiency almost untouched. Some non-canonical base pairing arises from incorrect polymerase activity or other mutagens (Blackburn and Gait 2006).

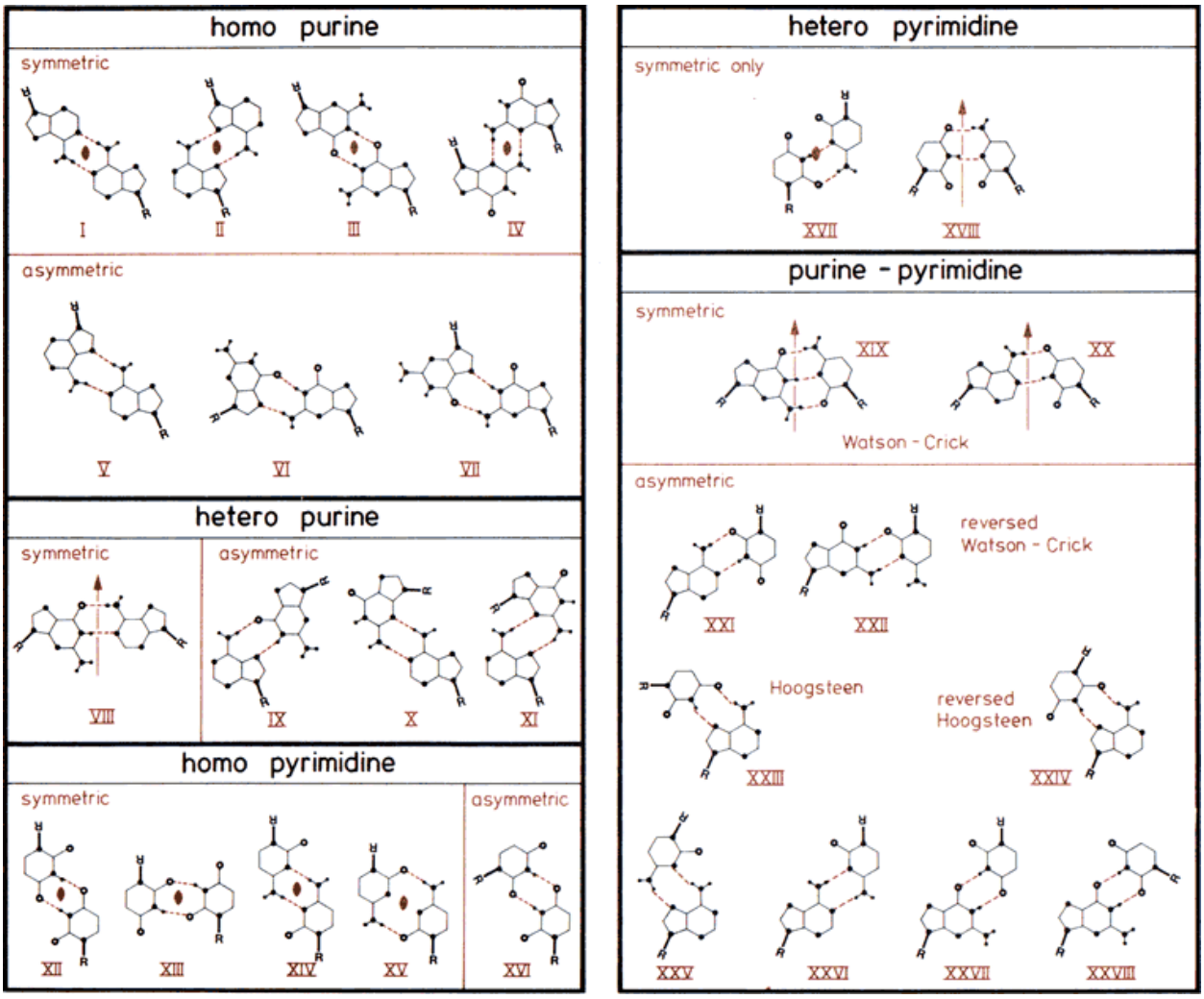

Figure 8: Notation of base pairing according to Saenger (Saenger 1984).

Architectures of DNA

Single-stranded DNA

Lesser known but for some biological processes essential form is single-stranded DNA (ssDNA). It is usually described to be a sort of an intermediate step during biological processes such as replication, recombination or transposition. The scarcity of this form is mostly caused by large hydrophobic patches around bases. It is therefore energetically convenient to bury hydrophobic surfaces. In other words ssDNA is thermodynamically unfavorable. In the living systems, ssDNA is readily bound to ssDNA-binding proteins (Blackburn and Gait 2006).

The information about their structure and function is limited. However, some information can be pieced together. Latest conformational view on ssDNA is that it is usually composed of stacked domains linked by random coil-like segments. Stacking effect must be taken into account for each

base step because it can differ significantly. A-rich tracts exhibit a stronger stacking effects compared to heteropolynucleotide, stacking effects become notable at lower temperatures. Even random sequences with no predicted higher structure will eventually show local hairpin, bulge or pseudoknot formation. Base pairing therefore cannot be avoided altogether but it can be predicted by available algorithms with a various degree of success (Liang, Kuhn et al. 2006).

Single-stranded DNA with enzymatic activity was reported, e.g., DNA metalloenzyme with nuclease activity (Cuenoud and Szostak 1995) or deoxyribozyme ligase (Silverman 2004). Understanding of folding patterns that outline formation of specific three-dimensional active molecules is currently a challenging task due to insufficient experimental data and limitations of theoretical approaches (Jeddi and Saiz 2017).

Helical forms of DNA

Nucleic acids are inherently flexible and in many cases highly polymorphic molecules. Our knowledge of nucleic acid architectures has grown significantly since the first helical models proposed by Watson and Crick. There are some architectural types that are found only in structures with strictly defined sequential requirements, some display relatively low or no sequential dependence (Neidle 2008).

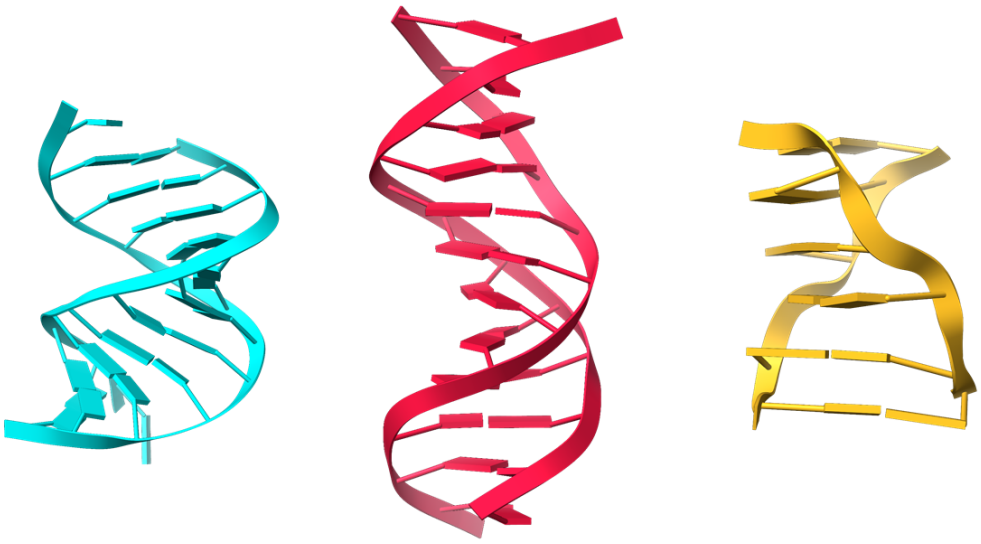

Figure 9: A-form (PDB 1zje), B-form (PDB 1bna) and Z-form (PDB 3p4j) of DNA ((Drew, Wing et al. 1981, Dohm, Hsu et al. 2005, Brzezinski, Brzuszkiewicz et al. 2011). Figure was created in ChimeraX (version 0.92) (Pettersen, Goddard et al. 2021).

It was clear from the initial structural studies that there are more helical forms (Figure 9), in which the difference can be quantified using helical and base pair parameters. More so, helical forms are interconvertible between each other. When DNA is present in relatively low humid conditions, the

A-form is preferred. Increase in humidity leads to changes in some helical parameters resulting in conversion to the B-form of DNA (Franklin and Gosling 1953, Neidle 2008).

B-form

The B-form of DNA, the classic form which was first modeled by Watson and Crick, was characterized by crystal structure studies on so-called Dickerson-Drew dodecamers, d(CGCGAATTCGCG) and various decamer structures in the eighties and nineties. These studies further refined general structural features obtained from fiber diffraction. Helix, assembled by two antiparallel strands, is right-handed and has 10 base pairs per complete turn. In the original Dickerson-Drew structure (Lawson, Artymiuk et al.), the helix is not straight but bent by about 19° in the major groove direction. Geometry of bases relative to helical axis results in lesser exposure of the hydrophobic surface, compared to the A-form. It was argued that this is the main reason why this form is preferred in a high humidity environment while the A-form is more stable due to its "economy of hydration" (Saenger, Hunter et al.).

From the base pairs point of view, their planes are perpendicular to the helical axis. Major and minor grooves are of similar depths, but differ in width. Bases are stacked almost exactly above preceding ones on the same strand. The sugars prefer the C2'-endo pucker and all glycosidic bonds are in anti-conformation or in so-called "high-anti" with χ values up to 240° (Neidle 2008).

As mentioned above, high humidity favors formation of the B-form. In better resolved crystal structures ordered water molecules can be seen in both grooves and around the phosphates. Major groove is filled with water molecules that can interact with carbonyl oxygens, amino- and inner ring nitrogens (Blackburn and Gait 2006).

A-form

As mentioned above, the A-form is induced in conditions of low humidity or in solution with alcohol. It strongly prefers C/G rich sequences, especially those with repeating Cs or Gs. As the B-form is the most stable conformation for DNA, the A-form is the most stable and prevalent in RNA where it forms the main scaffold of the molecular architecture. Most of the information about the A-form DNA duplexes comes from X-ray studies of octanucleotides (Neidle 2008).

This form is characterized by a wider right-handed helix (26 Å). There are also 11 bases per turn (28 Å). Base pairs are tilted with respect to the helical axis and the base centers are displaced from the helical axis. Glycosidic bond is in the anti-conformation, similar to the B-form but the χ torsion angle acquires lower values, close to 200°. Sugars occur in the C3'-endo pucker region. This results in significantly different groove characteristics when compared to the B-form - the major groove is deep and narrow and minor groove is wide and shallow. Displacement of the bases has also caused a hollow core along the helical axis when viewed from the top. Some sequences which

crystallized in the A-form may have the B-form in solution, as confirmed by NMR studies (Neidle 2008).

Z-form

Rampantly evolving technologies for quick and affordable oligonucleotide synthesis led to more freedom in designing sequences for X-ray experiments. In 1979, hexamer d(CGCGCG) was successfully crystallized and its structure was solved. Surprisingly, it turned to be a left-handed helix, now termed as the Z-DNA (Wang, Quigley et al. 1979).

There are 12 base pairs per helical turn, while helix being somewhat slimmer than that of the B-form. Guanine bases are in the syn conformation, cytosines in the anti. Consequence of this asymmetrical conformations are the geometrical differences between CG and GC steps. Edges of bases are on the surface of the helix rendering major groove almost non-existent. Minor one is narrow and deep (Neidle 2008).

Further crystallization experiments showed that change in the central region from C-G to A-T base pair is still tolerated from the point of view of maintaining the overall Z-DNA form. On the other hand incorporation of the A-T base pair disrupted an ordered hydration located in the grooves. This confirmed that the Z-DNA cannot be formed with sequences being composed solely of A-T base pairs (Blackburn and Gait 2006).

Summary of helical parameters of the main three DNA helical families are in the Table 1.

Table 1: Summarizes the structural features and parameters of the main three forms (Blackburn and Gait 2006).

	Base pairs per turn	Vertical rise per base pair (Å)	Helical twist (°)	Major groove - width (Å)	Major groove - depth (Å)	Minor groove - width (Å)	Minor groove - depth (Å)
A-DNA	11	2,54	32,7	2,2	13	11,1	2,6
B-DNA	10	3,38	36	11,6	8,5	6	8,2
Z-DNA	6	7,25	-60	8,8	3,7	2	13,8

Hairpin

Hairpin or stem-loop architecture is an intramolecular pattern, which is abundantly found in RNA and in some cases in single-stranded DNA. If two regions of the same strand are self-complementary, they can fold onto themselves creating a hairpin. Between two complementary parts there is usually some number of nucleotides that upon folding end up as an unpaired loop.

One such architecture is shown in Figure 10. This form of DNA often serves as an active interacting platform to enzymes such as nucleases. Hairpins can be exploited as probes for genomic detection or conductors for self-assembly of nanostructures (Wang, Dong et al. 2018).

Figure 10: Stem-loop solution model (PDB 1qe7) (Ghosh, Kumar et al. 1999). Figure was created in ChimeraX (version 0.92) (Pettersen, Goddard et al. 2021).

Stability of hairpin is determined by the base content in the stem and its length. The stem can be destabilized by bulges or mismatches. The length of the loop plays a significant role as well. Loops that are shorter than three nucleotides are sterically unfavorable while loops that are longer than eight nucleotides exhibit too much thermal movement. It has been proven that short hairpin (meaning short stem, 2-5 base pairs) depends on their loop residue and closing base pairs more than the longer variants. The thermal stability is usually tested by high resolution melting experiments (Rentzeperis, Alessi et al. 1993).

Multistranded DNA architectures

DNA can fold into an assortment of non-duplex structural architectures. They utilize non-canonical base pairing and feature a wide variety of complex topologies. Although it might seem like a rare occasion, they are frequently employed in many regulatory and structural functions (Neidle 2008).

Triplex

From Figure 11 it is apparent that when more strands are properly oriented to each other multiplexed structure can be formed. Specifically the third oligonucleotide can bind to the major groove of a B-form helix via the Hoogsteen edges. Resulting hydrogen bonding arrangement is called a triplet. From 1957, triple helix were continuously observed for a number of oligonucleotides. However, triple helices are less stable than duplexes due to presence of negatively charged phosphate groups on three strands. Triple helices are usually right-handed, third strand is typically bound in the major groove of the original duplex (Frank-Kamenetskii and Mirkin 1995, Rhee, Han et al. 1999, Neidle 2008).

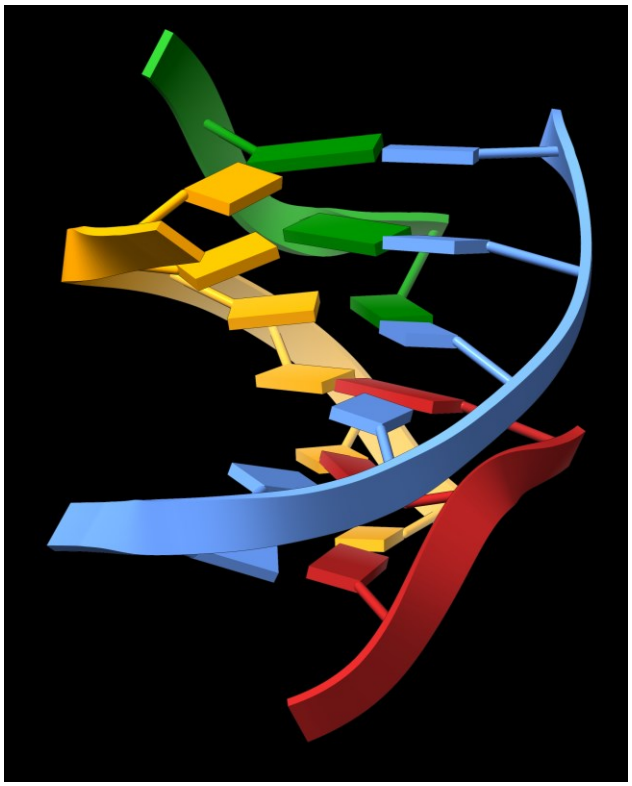

Figure 11: Triplex part of self-assembled four stranded DNA structure (PDB 1d3r) (Rhee, Han et al. 1999). Figure was created in ChimeraX (version 0.92) (Pettersen, Goddard et al. 2021).

Formation of the triplex can lead to an inhibition of transcription via binding of the third strand in the promoter region. Use of triplex in this way is unfortunately limited by suitability of promoter sequence (Blackburn and Gait 2006).

Guanine tetraplexes and i-motif

More prominent group of multiplexed structures are tetraplexes. When four guanines are bonded in a plane, it is termed the tetrade - fundamental building block of G-tetraplexes (Figure

12). These four stranded structural motifs can be found in guanine rich sequences. Guanine-rich sequences can form tetraplexes under physiologically relevant conditions and can be found at the telomeric regions of chromosomes. Guanines in tetrade are almost perfectly coplanar. In the middle of such tetrade is usually a stabilizing cation interacting with O6 oxygen in guanine residue. Some cations, such as K^+, can be positioned between two tetrads (Neidle 2008).

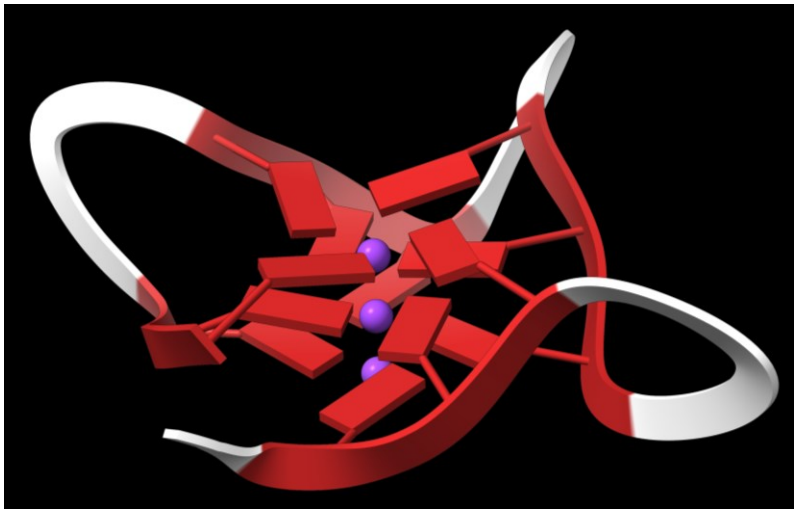

Figure 12: G-tetraplex of telomeric sequence (PDB 6ip3), nucleotides in the loops are not shown for the sake of clarity, K^+ cations are depicted as purple spheres (Nuthanakanti, Ahmed et al. 2019). Figure was created in ChimeraX (version 0.92) (Pettersen, Goddard et al. 2021).

G-tetraplexes. More strands involved in formation of G-tetraplexes naturally lead to far more complex topologies than in the case of two stranded helices. G-tetraplex can be intermolecular (formed by more strands) or intramolecular (formed by one strand), these two ways of folding differ in the orientation of the backbone. Structures tend to display pseudo helical twist of 36 º (Neidle 2008).

G-tetraplexes can significantly differ in topology, even one sequence can often take more than one topology in solution. Topology of G-tetraplexes can be classified using a polarity of the strand - parallel or antiparallel. Each group can be divided according to the character of the loop, if present. Loop can be lateral, diagonal or propeller as shown in Figure 13 (Burge, Parkinson et al. 2006).

All tetraplexes have four grooves, their dimensions are type-dependent. G-tetraplexes with a propeller twist show more complex groove characteristics opposite to the relatively straightforward “all lateral” or “all diagonal” loop topologies. All above mentioned structural variables (length of the sequence, loop topology, strand polarity etc.) contribute to the vast conformational variety. Prediction of 3D models based on the sequences has been researched extensively but only general trends have been found, e. g. longer sequences likely prefer lateral and diagonal loops etc. (Burge, Parkinson et al. 2006).

Most studies were done on telomeric repeats and on sequences derived from promoter regions (such as MYC and KIT gene promoter). Some X-ray structures have been solved to an exceptional resolution of 0.95 Å and thus giving a profound insight to the structural features. Application of machine learning on large structural and sequential datasets has been adapted recently, further research courses include effects of molecular crowding and base modifications on the stability of tetraplex structures (Spiegel, Adhikari et al. 2020).

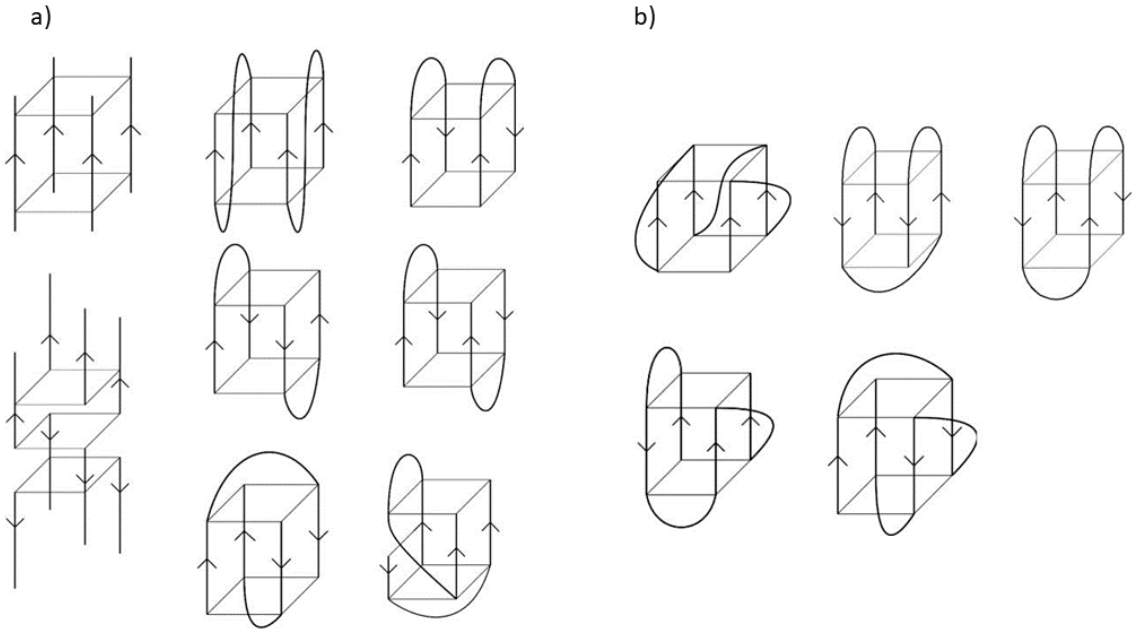

Figure 13: Tetraplex loops and topology, a) show various topologies for tetrameric and dimeric G-tetraplexes and b) for unimolecular ones. Arrows indicate directions of each strand involved in the topology (Burge, Parkinson et al. 2006).

Detection of G-tetraplexes exploits binding of either small molecule, usually on top of structure followed by fluorescent probing, or protein (antibody). The next-generation sequencing techniques have been adapted as well (Spiegel, Adhikari et al. 2020).

The i-motif. Crystals grown in acidic conditions with cytosine rich sequences yielded intercalated tetraplexes, Figure 14. Acidic pH plays a crucial part because it allows one of the cytosines to be protonated on the N3 nitrogen and therefore enables a hemi-protonated C-C^+ base pair to form *in vitro* conditions. Their biological role was disputed because of the necessary protonation in acidic pH values, however cell environment can compensate this via local pH fluctuations and molecular crowding effect. Evidence for this is recent observation that i-motif can form in human nuclei, where they have been detected with selective antibody and it is cell cycle dependent (Blackburn and Gait 2006, Neidle 2008).

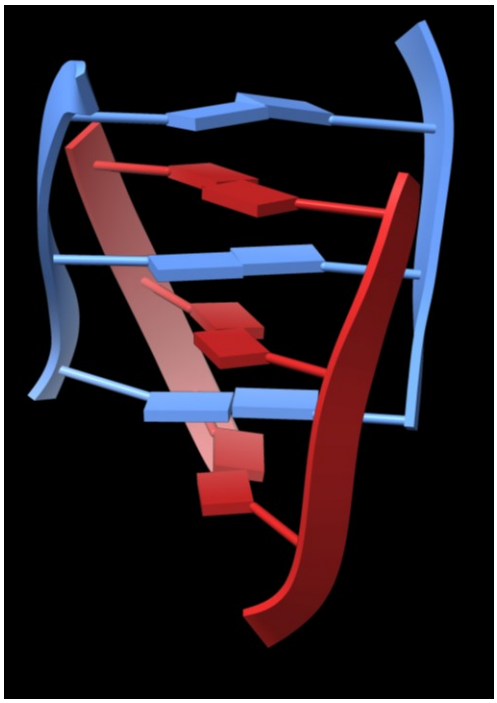

Figure 14: Intercalated cytosine stretches of i-motif structure (PDB 1cn0) (Weil, Min et al. 1999). Figure was created in ChimeraX (version 0.92) (Pettersen, Goddard et al. 2021).

Other studies compared cytosine rich DNA sequences in order to determine their stability under near physiological conditions. It has been shown that at least five consecutive cytosines are required for folding stable i-motif under room temperature and near neutral pH. Negative superhelicity, low temperature and presence of specific cations favor formation of i-motif greatly. It is suggested that i-motif could be a potential target for therapy given they are complementary to guanine rich sequences (Abou Assi, Garavís et al. 2018).

Experimental methods for studying nucleic acids

Structural study of nucleic acids are historically connected with diffraction techniques. Use of nuclear magnetic resonance has proved to be valuable when analyzing dynamic properties of nucleic acids in solution. Other methods providing lower resolution include but are not limited to circular dichroism (CD), microcalorimetry, FRET, small angle X-ray scattering (SAXS) and last but not least cryo-electron microscopy. Of these biophysical methods, only basic principles of X-ray crystallography, nuclear magnetic resonance, circular dichroism, and SAXS are mentioned here explicitly (Nelson 2005, Blackburn and Gait 2006, Neidle 2008).

X-ray crystallography

To this day it is the most dominantly used method researchers rely on when structure on atomic scale is needed. Quick look at deposition data at PDB confirms it. Diffraction on the monocrystal is preferentially used over outdated fiber diffraction nowadays. This method is based on the interaction of macromolecules in an ordered crystalline phase with the X-ray beam. This section will give an overview of a process that starts with a obtaining of biological macromolecules and ends with solved structure on the atomic scale. There are several critical steps, first of all one must produce and purify biomacromolecules in sufficient amounts and best possible purity (Neidle 2008).

Purification of the macromolecules

Chemical synthesis of nucleic acids based on the solid phase method is a viable choice when short oligonucleotides are desired. It is not recommended practice to acquire strands longer than 100 nucleotides in this way as well as strands with repeats longer than 10 nucleotides. For large complexes such as ribosomes, an approach similar to protein purification must be employed, briefly described in following paragraph. Sometimes, only biologically interesting regions are crystallized while nonessential regions are omitted or replaced with motifs that promote crystallization, most likely folds and their sequential dependencies may be estimated *in silico* (Mooers 2008).

Purification of proteins is to some extent an empirical procedure, in which for almost every protein of interest researchers choose a specific combination of expression system and purification steps. Expression systems vary from bacterial (strains of *Escherichia coli* such as TOP10 or B21) or yeast (*Saccharomyces cerevisiae* EBY100 strain) to insect cells (Schneider 2 cells) and mammalian cultures (HeLa or HEK-293). Cell free expression has been utilized recently as well. It is based on *in vitro* reaction of all essential elements needed for protein production (RNA polymerase, regulatory factors, transcription factors, ribosomes, DNA template etc.) In order to

successfully obtain monocrystals, it is recommended that concentration of protein or DNA should be at least 5-10 mg/ml (Rhodes 2006, Mooers 2008).

Crystallization

Growth of crystals can be best described with a phase diagram, Figure 15. The basic principle rests in the organized phase transition of macromolecules of interest from liquid, solution phase to solid, crystal phase, which can be simplified as three dimensional ordered array of macromolecules. Space between them is filled with semi-ordered solvent molecules and disordered solvent molecules in the bulk solvent, it is stated that bulk solvent is responsible for 50 or in some cases even more percent of total crystal volume (Biedermannova and Schneider 2016).

The most common method to prepare crystals of proteins and nucleic acids is based on controlled evaporation of water to raise concentration of macromolecule and precipitant. Crystals should appear when concentration of macromolecule and precipitant exceeds certain value. There are two steps in the crystallization process - nucleation and growth. The large supersaturation is required in order to overcome the free energy barrier which is present when forming the nucleus - microscopic array of macromolecules. At the point when there is a sufficient amount of nuclei it is desired to move the system to a zone where no other nucleation is supported but crystal growth can occur, so called metastable zone (Ducruix and Giege 1992, Neidle 2008).

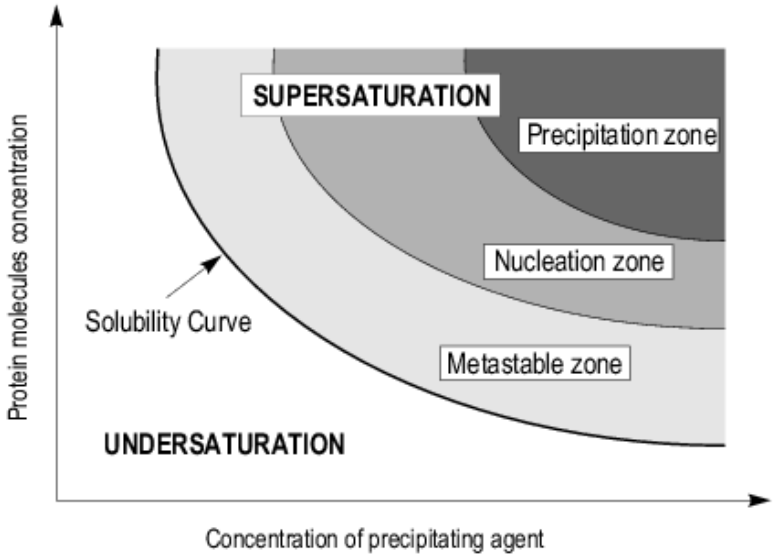

Figure 15: Phase diagram describing nucleation and following crystallization (Ducruix and Giege 1992)

There are various possible setups that take advantage of the principle described above. The most frequently used are vapor diffusion methods called hanging drop and sitting drop. This technique allows macromolecule/precipitant mixture to equilibrate with significantly larger reservoirs in

closed containers. In the hanging drop setup, a few microliters of macromolecule mixture and reservoir solution are mixed in desired ratio onto glass cover slip. Once the container is sealed, water molecules can transfer from drop to the reservoir until the system reaches equilibrium (Rhodes 2006).

Protein and nucleic acid crystals are much more fragile than crystals of organic or even inorganic compounds. Very gentle methods and techniques are therefore required when handling them. Crystallization of nucleic acids has a couple of unique challenges compared with protein crystallography. In contrast to protein molecules which have a variety of chemical and structural groups on the surface which enable crystal contacts, the surface of nucleic acids is dominantly composed of negatively charged phosphate groups which can, in turn, lead to difficulties with crystal packing. This, however, can be reduced using appropriate buffer solutions with higher concentrations of cations such as: Mg^{2+}, Na^{+}, K^{+} or NH_4^{+} or simple organic amines such as spermine or spermidine (Mooers 2008).

Whether the macromolecule crystallizes or precipitates to amorphous solid is dependent on many properties other than macromolecule and precipitant concentrations. Temperature, pH, ionic strength, concentration of additives and even tiny vibrations might have an effect on successful crystallization. Chain length of nucleic acid should be considered as well. Some macromolecules can be inherently unstructured and different approaches must be considered (Rhodes 2006, Mooers 2008, Nanev 2020).

Diffraction experiment

Crystal is mounted between X-ray source and detector. Sources of the X-ray beam for macromolecular crystallization are in most cases local diffractometer with the rotating anode, newly metal-jet type anode or large synchrotron facilities (Rhodes 2006).

Detector detects positions and intensities of the diffraction spots, also called reflections, but only those that obey the geometric principle of diffraction, described by Laue equations or Bragg condition, are detected. They are based on the constructive interference of X-ray. Position and intensity of each reflection contains information about real lattice via inverse relationship with reciprocal lattice. If a crystal is rotated in the X-ray beam, different images are detected corresponding to the cross section of the reciprocal space and the Ewald sphere. Reflections outside of this sphere cannot in principle be detected. Diffracted beam is not linear but rather cone-like shaped. That is the reason why reflections are not detected as points but they are spherically shaped, the consequence of non-ideality of measured crystals. The quality of crystal is judged by their ability to produce the sharpest reflections at as high resolution as possible. That is checked by taking a few preliminary diffraction images, if they are sharp enough the entire data set can be collected (Rhodes 2006).

Macromolecular crystals, composed of flexible large molecules held by weak interaction, show greater mosaicity than crystals of small molecules. The reflections therefore suffer from bigger mosaic spread (high mosaicity). This can be of some use because the spots (reflections) are broader and therefore measurable. This is true only to the point when the mosaicity is too high causing the reflection spots to overlap (Rhodes 2006).

Crystallographers assign Miller indices to each reflection. These integer numbers are denoted as h, k and l. The central reflection, while experimentally unobtainable, has coordinates (h, k, l) = 0, 0, 0, or hkl = 000. Other reflections are assigned integer numbers. The diffraction experiment is described by a series of numbers hkl, and measured reflection intensities I_{hkl}. The most outer measurable reflection on the diffraction image, the reflection with the highest crystallographic resolution, is said to give potential resolution of the model we can obtain. The "reciprocal space" of Miller indices hkl and intensities can be transformed by Fourier synthesis into the "real space" of atomic positions described by x, y, z coordinates (Rhodes 2006).

Structure factor and electron density map

Reflected rays are treated as waves which can be recombined to produce the content of the unit cell. The resulting wave is quite sophisticated because each reflection is a combination of diffraction from rather intricate objects such as macromolecules. Every wave can be mathematically expressed in the terms of periodic functions sines or cosines (example Eq. 1)

$$f(x) = F\ sin\ 2\pi\ (hx + \ \alpha). \tag{1}$$

Where F is the amplitude of the wave, h specifies the frequency and α is the phase of the wave. The phase information can be mathematically understood as the position of the entire wave with respect to the origin of the plot. Equation 1 describes a one dimensional wave. Complicated periodic function can be expressed as a sum of periodic functions, Eq. 2. It can then be rewritten using complex numbers, Eq. 3

$$f(x) = \sum_{h=0}^{n} F_h \sin 2\pi\ (hx + \alpha_h) \tag{2}$$

or

$$f(x) = \sum_{h} F_h e^{2\pi i(hx)} \tag{3}$$

In three dimensional space we can than state following equation (Eq. 4):

$$f(x,y,z) = \sum_x \sum_y \sum_z F_{hkl} e^{2\pi(hx+ky+lz)} \tag{4}$$

Each reflection produced from a diffracted X-ray beam can be described as the sum of the contributions from all diffracting elements in the unit cell. The structure factor equation is the sum that characterizes the diffracted X-ray. The sum for reflection hkl is the structure factor F_{hkl}. The structure factor equation can be viewed as the sum of terms, where each term describes diffraction by one atom in the unit cell. The structure factor F_{hkl} is a Fourier sum (Rhodes 2006).

Diffraction reveals the distribution of electrons in the unit cell since the actual diffractors are electrons. Electron density thus reveals the shape of molecules. We can take advantage of the fact that molecules are in the form of an ordered array in crystals, thus the electron density is from a mathematical point of view a complicated periodic function, ρ(x, y, z). Graph of this function is an electron density map. In essence, the goal of crystallography is to gain the function whose graph is the electron density map (Rhodes 2006).

F_{hkl} can be rewritten as the sum of contributions from each volume element in the unit cell. If we make the volume element smaller, more precise average electron density we get in all points of the map. For infinitesimally small volume elements, we can write equation (5):

$$F_{hkl} = \int_x \int_y \int_z \rho(x,y,z) e^{2\pi i(hx+ky+lz)} dx\, dy\, dz \tag{5}$$

Each reflection is described by corresponding structure factor equation Fhkl, giving us a large number of equations for the ρ(x, y, z) function. Operation called Fourier transform solves the structure factor equations for the desired ρ(x, y, z) function. The Fourier transform precisely describes the relationship between objects in the unit cell and their diffraction pattern. Therefore if we have three information (amplitude, frequency and phase) for each reflection, we can obtain the desired ρ(x, y, z) function (Rhodes 2006).

Unfortunately only intensity I_{hkl} and position of reflection is experimentally accessible. Intensity provides information about amplitude of the wave (it is proportional to F_{hkl}^2) and position of the reflection gives information about the frequency. Phase alpha for any reflection cannot be measured on any kind of detector (Rhodes 2006).

Phase problem

The inability to obtain information about the phase of each wave is called the phase problem. Fortunately, there are certain strategies that can aid crystallographers to achieve successful determination of the structure. It is useful to know that phase carries a far more information than directly measurable intensities (Rhodes 2006). One can look at famous duck and cat images by Dr David Cowtan to appreciate how the correct phase can affect the final image.

If we add a small number of atoms to identical sites in each unit cell in the crystal we would see discernible changes in the diffraction pattern. This shift has a root in the fact that the introduced atom influences all reflections, some weakly but those reflections which are related to the lattice planes that intersect directly with that atom are influenced strongly. This method for obtaining phases is called isomorphous replacement. It is obvious that the atom must be a strong diffractor, therefore must have considerably more electrons. This condition is checked for heavy atoms such as Hg, Pb or Au etc. (Rhodes 2006).

Introduction of heavy atoms can be achieved via soaking of the crystal in heavy atom rich solution (Selenium in solution of selenourea) or it can be introduced during synthesis and be covalently linked in the macromolecule (bromouracil or iodouracil substitutes thymine residue in nucleic acid chain). When a derivative crystal is obtained it must be isomorphous with the native one, it cannot disrupt crystal packing and therefore dimensions and symmetry of the unit cell. Derivative crystals should diffract to a reasonably high resolution, although not necessarily to resolution of the native data (Rhodes 2006).

We are able to directly find the position of the heavy atom in the unit cell with the use of Patterson synthesis, therefore we know **F**(H) including its phase angle. To find **F**(M) we place -**F**(H) in the origin of the complex plane and draw a circle (radius of |**F**(MH)|) on the end of vector -**F**(H). Head of the **F**(MH) vector lies on the circle. Next at the origin we draw a circle (radius of **F**(M)) and thus reveal two points where circles intersect. This projection is called Harker diagram, Figure 16. Relationship between **F**(HM), **F**(H) and **F**(M) is described by equation 6. It is common practice to use a second heavy atom derivative which to some extent confirms which phase is correct (Rhodes 2006).

$$\boldsymbol{F}(HM) = \boldsymbol{F}(H) + \boldsymbol{F}(M) \tag{6}$$

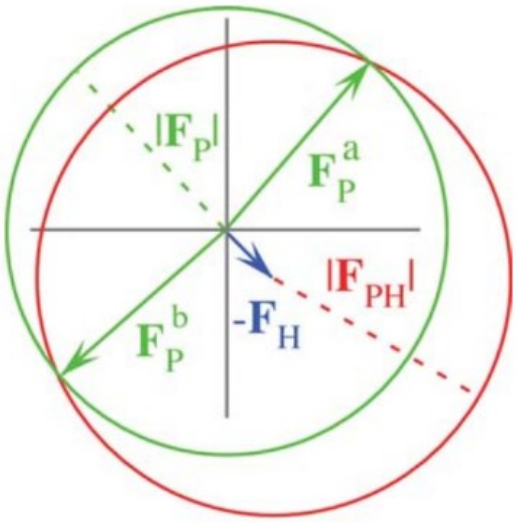

*Figure 16: Harker construction for estimating phase for protein structure factors of protein crystal **F**(P) using heavy-atom derivative data set **F**(PH) (Rhodes 2006).*

Anomalous scattering is the second option to obtain phases. It takes advantage of disruption of intensities in the symmetry-related reflections due to absorption of specific wavelengths by heavy atoms. Every element has a unique absorption wavelength, just below emission wavelength kβ, at which the absorption drops. When plotted, this absorption difference as a function of wavelength is called absorption edge. It is said that the element exhibits anomalous scattering when X-ray wavelength is near the absorption edge. This technique therefore requires tunable wavelengths of X-rays, so synchrotron sources are utilized (Rhodes 2006).

When above conditions are met, the structure factor **F**(MH) is influenced by two contributions from heavy atom - real and imaginary (Eq. 7).

$$\boldsymbol{F}(MH)_2 = \boldsymbol{F}(MH)_1 + \boldsymbol{F}(r) + \boldsymbol{F}(i) \tag{7}$$

F$(MH)_1$ and **F**$(MH)_2$ are structure factors measured at two different wavelengths. Magnitudes of **F**(r) and **F**(i) for each element can be looked up. They only depend on the position of the atom in the unit cell, which can be determined using Patterson method similarly as in the case of isomorphous replacement method. Full knowledge of the heavy atom contributions can solve equation above for **F**(MH)1, resulting in phase (Rhodes 2006).

It is common to combine both methods called SIRAS. The process consists of collecting amplitudes for native crystal |FM|. After that heavy-atom derivative dataset is collected, giving amplitudes of |FMH|. Finally, the third dataset is collected, but now at a different wavelength. We use third set and non-equivalence of Friedel pairs to get phases from heavy atom derivative and then the phased heavy atom derivative structure to obtain the native phases (Rhodes 2006).

Molecular replacement can sometimes be used when phases from structure factors of known macromolecule structure (phasing model) are available (Rhodes 2006).

Building a model and refinement

Once the graph of ρ(x, y, z) is obtained, we interpret it by building a model into it. Prior information about macromolecules are used at this point. The final model must be consistent with chemical aspects (length of bonds, conformational angles etc.) (Rhodes 2006, Mooers 2008).

Improving electron density maps and models is an iterative process. It is necessary because phases are usually coarse estimates, datasets from derivative crystals are often at lower resolution thus first maps may be inaccurate. Each set of phases has a reliability factor (figure of merit) which is used as a weighting factor for Fourier computation of the map, Equation 8. This process ensures that terms with low reliability have a reduced contribution to the Fourier sum (Rhodes 2006).

$$\rho(x,y,z) = \frac{1}{V}\sum_{x}\sum_{y}\sum_{z} w_{hkl}|\boldsymbol{F}_{obs}|e^{-2\pi i(hx+ky+lz-\alpha'_{calc})} \tag{8}$$

First map now serves as a model of structure, at this point it is used to improve by tuning the function (density modification) so it depicts macromolecule with as much accuracy as possible. Next, this modified map is assigned a low value of ρ(x, y, z) where the bulk solvent is estimated and high value for regions where macromolecule is located. Each point of the map is then analyzed for the value of the ρ(x, y, z) function, where ρ(x, y, z) is negative it is assigned a zero. If the value of ρ(x, y, z) is positive, it is averaged within a defined distance. Result of this action is a smoother map and is now used to calculate new structure factors. They should reveal values of amplitudes and phases if the initial model is correct. Newly obtained phases are again used with $|F_{obs}|$ to determine ρ(x,y,z) (Rhodes 2006).

If the new phases improve the density map, it will be more detailed so it should define molecular boundaries better and the process is repeated. Each successive map should be clearer and more precise. Eventually phase estimates might converge beyond the heavy atom derivative. This is the base for a process called phase extension, in which phase assignments are extended to a higher resolution because improved phase estimate improves resolution of the map (Rhodes 2006, Mooers 2008).

At some point in the refinement, the map becomes clear enough to fit a macromolecule in it. If so, we can begin constructing a molecular model. Similar procedure as described in the previous article is employed again. From the model we calculate the structure factors, additional iterations improve the map further which allows more molecular details to be introduced. However, conversion to a molecular model potentially increases bias from model into electron density

function. To avoid extreme cases where the series is composed from amplitudes purely from intensity data and phases purely from model, additional Fourier calculations of the map can decrease such a bias. This results in a map called F_{obs}-F_{calc} or F_o-F_c. Depending on which of those two components has a larger value we can interpret the map. Negative density implies that the model imposes more electron density that the unit cell contains and we should move atoms away from such a region. For example, wrong conformation of amino-acid or base residue might show negative density peak and positive peak nearby could point to correct one. $2F_o$-F_c map is positive almost everywhere, exceptions are regions with severe errors. It can be read as a difference F_o-F_c map with electron density around the macromolecular model (Rhodes 2006).

Last stages of structure determination are dominated by altering of reciprocal space refinement and with map fitting (real-space refinement). Specialized version of least-square approach was historically used in the refinement but more sophisticated methods are utilized nowadays, for example Bayesian statistics. Modern programs are capable of automated refinement cycles starting with random distribution of atoms in electron density and proceed to models with only some residue positions that require manual building (Rhodes 2006).

Evaluation of refinement process

R-factor is one of the most widely used measures of convergence measured F_{obs} and calculated F_{calc}. It is defined by Equation 9 as:

$$R = \frac{\sum\left||\boldsymbol{F}_{obs}| - |\boldsymbol{F}_{calc}|\right|}{\sum|\boldsymbol{F}_{obs}|} \tag{9}$$

R-factor values are naturally in the range from 0 to about 0.6. Value is approximately 0.6 is achieved when F_{obs} is compared with a set of random structural factors. However, the value of R beyond 0.5 is considered as very poor. Early attempts with R-factor around 0.4 are promising and the final value for large macromolecules should be around 0.2. To put those values in the context, small organic molecules can be refined to R-factor below 0.1 (Rhodes 2006, Nanev 2020).

Free R-factor, R_{Free}, can be computed with a small set of intensities (randomly chosen) set aside during refinement. The aim is to test how well the current model can predict those missing intensities. During iterations, R_{Free} values are higher than R, but in the final stages they should almost coincide (Rhodes 2006, Nanev 2020).

Nuclear magnetic resonance

Detection of nuclei with nuclear spin in an external magnetic field is the principle of nucleic magnetic resonance (NMR). The interaction of nuclear magnetic moment with magnetic field causes a splitting in the energy of the spin states, for nuclei with spin ½ there are two energy states:

higher and lower state. At thermal equilibrium, there is a slight excess of nuclei in the lower energy state but when radiofrequency pulse, which matches the energy gap between two states (resonance condition), is applied upon them, transition between the nuclei states population occurs. Since the energy gap is characteristic for every nucleus and each one is matched with specific frequency, called Larmor frequency, it can be used for structure determination purposes (Neidle 2008, Al-Hashimi 2013).

Nuclei suitable for NMR experiments are present in hydrogen atoms which are abundant in biologically significant macromolecules. ^{13}C and ^{15}N labeled oligonucleotides are now readily available and their incorporation has enabled expansion of NMR structural as well as dynamic techniques. Absorbed energy (detected signal) is proportional to the difference in the state population. Term chemical shift is used to describe the position of each NMR signal corresponding to active nuclei. However detected signals are dependent on the shielding effects (electronic structure) of neighboring atoms, mostly protons (Al-Hashimi 2013).

NMR is a spectral technique and thus it gives us a series of indirect information and only when they are properly interpreted can elucidate structure and dynamics of the examined system. Spectroscopic studies could describe base-pairing pattern, site specific interaction between nucleic acid and ligands etc. (Neidle 2008).

First step is assignment of active nuclei to its resonance in the NMR spectra. After that NMR spectra are interpreted in terms of NOE contacts, J couplings and cross-correlated relaxation rates for acquiring a 3D model. Compared to assignment for proteins, nucleic acids are more complex due to having only four major components, therefore chemical shift dispersion is reduced. Similarity of the chemical environment of nucleotides in the dominant helical form is the main reason for similar chemical shifts. On the other hand shift dispersion is observed in non-canonical elements, which makes NMR valuable for RNA studies. Nature of the helical forms generally does not permit the presence of long-range correlation (Zidek, Stefl et al. 2001, Al-Hashimi 2013).

Imino proton resonance of the guanines and uracils in range 10 - 15 ppm holds information about base pairing in the RNA. By integrating under the spectral line, the number of such pairings can be obtained. Canonical pairing can be found in the region of 12 - 15 ppm. Thus one dimensional spectrum presents insight into pairing patterns. Spectra measured in $^{2}H_2O$ is used to find non exchangeable protons on sugar and base moieties. Modeling is used to generate 3D representation of the molecule. Coupling constants and NOE derived distances are used to determine conformation. These distances are taken as constraints to molecular dynamics which is applied to the crude starting model (Adrian, Heddi et al. 2012, Al-Hashimi 2013).

Advantage of NMR compared to X-ray crystallography lies in the ability to perform dynamic studies in the almost native solution - liquid phase. NMR therefore emphasizes the flexible nature of nucleic acids and can map their dynamic behavior. Obvious limit is the size of the examined

system, if the molecules are too large the spectrum becomes far too complex to interpret (Blackburn and Gait 2006, Neidle 2008).

Circular dichroism spectroscopy

Different absorption of right-handed and left-handed circularly polarized light by chiral molecules is called circular dichroism (CD). In nucleic acids there are three sources from which CD signals can be detected. First, the asymmetric sugar, specifically C1' atoms, second is the inherent helicity of the polynucleotide chain and third comes from the long range intermolecular interactions in some environments. Although theory of CD spectroscopy is well-based the experimental use is still mostly empirical (Kypr, Kejnovska et al. 2009).

CD spectroscopy uses a spectral range of 200-320 nm. Measurements in the UV range are more sensitive and give more information but they are difficult to perform because they require specialized instruments while CD in the infrared region is less sensitive. CD spectroscopy can give insight into the overall topology of the sample and can examine polymorphic nature of nucleic acids (Vorlickova, Kejnovska et al. 2012).

Disordered or denatured DNA exhibit weak spectral features due to the lack of chirality. The B-form of the helix gives CD spectrum with positive band around 275 nm and a negative band at around 245 nm. Intensities of those two peaks are relatively the same and the spectrum is conservative, meaning the integral of the spectral curve is close to zero. The A-form of DNA and RNA give a much stronger CD spectrum than the B-form, which is probably caused by a tilting of the base pairs and resulting weaker stacking. Spectrum of the A-form is dominated by strong positive peaks around 260 nm and negative one at 210 nm. The CD spectrum of the Z-form mirrors the characteristics of the B-form spectrum (Vorlickova, Kejnovska et al. 2012).

CD spectroscopy can quickly and with relative precision distinguish topology of G-tetraplexes. The spectrum of parallel tetraplex is made of positive strong band around 260 nm while antiparallel is mainly composed of positive 295 nm band and negative one around 260 nm. The 260 band is usually similar in shape with the one found in the A-form spectrum. Both tetraplex topologies have a positive band around 210 nm (Vorlickova, Kejnovska et al. 2012) which is unfortunately hidden in absorption caused by most buffers.

Big advantage of this method is that the experiment is fast and relatively cheap. CD spectroscopy is sensitive and therefore requires a small amount of material. Variability of experimental condition (pH, temperature, titration with cations, etc.) is quite useful, because we can gain information on changes under certain conditions. It is important to always look at the whole spectrum when interpreting it. It has been proved that CD spectroscopy is a powerful method and can give complementary information to X-ray and NMR study (Vorlickova, Kejnovska et al. 2012).

SAXS

Small-angle X-ray scattering is a technique that allows to quantify differences of material density of the sample. Orientation averaged scattering pattern is obtained during the experiment. The sample is exposed to X-rays and the detector registers a scattered radiation. The experiment is performed very close to the primary beam. In the best case, the resolution obtained is about 15 Å. Therefore this method does not provide atomic-scale resolved structures, when compared to X-ray crystallography, but can contribute with conformational alternatives to such structures. Similarly to NMR or CD, experiments with biomolecules are performed in the aqueous solution (Blanchet and Svergun 2013).

The scattered intensity is recorded as a function of momentum transfer. Scattering of the buffer solution is subtracted before further processing. For a monodisperse particle solution, intensity distribution is averaged over all orientations. Data from intermediate to high scattering angles provide information about overall size and novel algorithms allow ab initio reconstructions from scattering profiles. The most common application of SAXS is to determine the radius of gyration which can be obtained from data at lowest scattering angles via Guinier fit from samples with low concentrations (intermolecular scattering is negligible). Additionally, molecularity or oligomeric state can be obtained as well (Dyer, Hammel et al. 2014).

Bioinformatic tools

NtC

Novel approaches to describe architectural features of both DNA and RNA can be achieved with the use of diNucleoTide Conformers, NtC. They describe the geometry of the dinucleotide step using nine torsions (δ1, ε1, ζ1, α2, β2, γ2, δ2, χ1 and χ2), pseudotorsion μ and distances N'N' and C'C' (Figure 17). With these parameters, NtC provides understanding of the structural behavior of the backbone reflecting its plasticity. Such information was not available before the NtC classes were defined (Cerny, Bozikova et al. 2020).

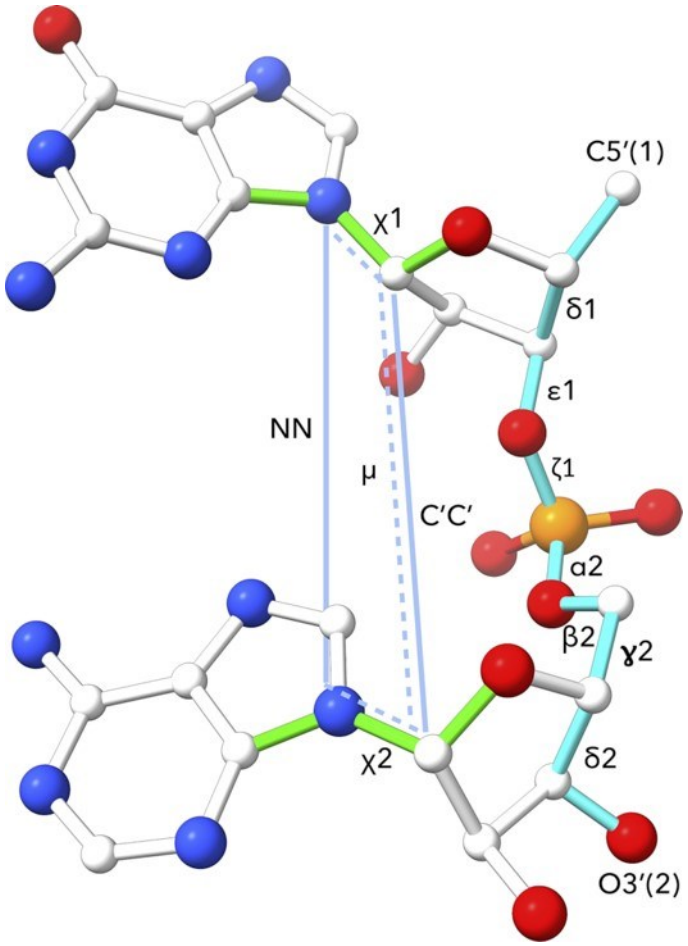

Figure 17: Definition of the parameters: δ1 C5'(1)–C4'(1)–C3'(1)–O3'(1), ϵ1 C4'(1)–C3'(1)–O3'(1)–P(2), ζ1 C3'(1)–O3'(1)–P(2)–O5'(2), α2 O3'(1)–P(2)–O5'(2)–C5'(2), β2 P(2)–O5'(2)–C5'(2)–C4'(2), ɣ2 O5'(2)–C5'(2)–C4'(2)–C3'(2), δ2 C5'(2)–C4'(2)–C3'(2)–O3'(2), χ1 O4'(1)–C1'(1)–N1/9(1)–C2/4(1), χ2 O4'(2)–C1'(2) N1/9(2)–C2/4(2), the parameters NN as N1/9(1)–N1/9(2), C'C' as C1'(1)–C1'(2) distances and pseudotorsion μ as N1/N9(1)–C1'(1)–C1'(2)–N1/N9(2) from (Cerny, Bozikova et al. 2020). Figure was created in ChimeraX (version 0.92) (Pettersen, Goddard et al. 2021).

The NtC classes were determined based on a set of sequentially nonredundant structures. Structures of their dinucleotides were submitted to clustering methods. Clusters were critically evaluated and some of them were merged due to geometrical closeness, this resulted in definition of 96 + 1 NtC classes. Golden set is a manually curated group of close to 7 000 dinucleotide steps that defines 96 NtC classes. Formally introduced class 97 is reserved for the unassigned steps (NANT). They have been further grouped into letters of the CANA alphabet (Conformational Alphabet of Nucleic Acids). The relationship between the detailed NtC classification and more intuitive CANA letters is shown in Table 2. The geometry and brief structural description is described on the dnatco.datmos.org website. Assignment of NtCs to structures can be done there as well (Černý, Božíková et al. 2020).

Table 2: CANA letters and corresponding NtCs.

CANA letter	NtC classes merged in CANA letter
AAA	*AA00 + AA02 + AA03 + AA04 + AA07 + AA08 + AA09 + AA12 + AA13*
AA1	*AA01 + AA05 +AA06 + AA10 + AA11*
A-B	*AB01 + AB02 + AB03 +AA04 + AB05*
B-A	*BA01 + BA05 + BA08 + BA09 + BA10 + BA13 + BA16 + BA17*
BBB	*BB00 + BB01*
BB1	*BB02 + BB03 + BB17*
B12	*BB04 - BB05*
BB2	*BB07 - BB08*
miB	*BB10 – BB16 + BB18 + BB20*
SYN	*AA1S + AB1S + AB2S + BBS1 + BB1S + BB2S*
ICL	*IC02–IC03 + IC05–IC07*
OPN	*OP01-OP22 + OP24*
ZZZ	*ZZ1S + ZZ2S + ZZS1 + ZZS2 + ZZ01*

The geometrical closeness of the analyzed step to NtC class is quantified with the confal score. It is calculated as the harmonic mean of the twelve parameters that define the geometry of dinucleotide steps. The assignment protocol can be roughly outlined in the following points. Firstly structure is uploaded and checked for presence of the nucleic acids, values of the twelve parameters are then calculated. After that, the distances between step and all the members of the golden set are determined. Analyzed step is assigned to the NtC class of the nearest neighbors (Cerny, Bozikova et al. 2020).

The power of NtCs lies in their ability to conveniently annotate nucleic acid structures with relative ease (Schneider, Bozikova et al. 2017). As it was demonstrated in recent studies, it opens the possibility to improve analyzed structures with the knowledge acquired from the assignment of NtCs (Schneider, Bozikova et al. 2017, Cerny, Bozikova et al. 2020).

The Protein Data Bank

The need for a unified and curated source of structural information was the motivation for founding of structural databases. The most used primary structure database is the Protein Data Bank (PDB) which was established in 1971 as a repository of biological crystal structures. PDB is now managed by wwPDB with participation of the Research Collaboratory for Structural Bioinformatics (RCSB), European PDBe, and Japanese PDBj. PDB database contains more than 170 000 entries of biological macromolecules and is regarded as a fundamental science resource (Burley, Berman et al. 2018).

The aim of any structural database is to annotate and organize data that contain information about structures such as spatial atomic coordinates, information about the experiment or bibliography among others. Each entry in the PDB database is assigned a unique four character long alphanumeric code called PDBid. The PDB allows users to search for entries directly by its PDBid or offers comprehensive searching methods. When desired structure or list of structures are found users can view 3D representation via some of the built-in molecular viewers, download custom or pre-formatted reports in the csv file or further improve the search query (Berman, Westbrook et al. 2000).

The information of each entry is stored in the Macromolecular Crystallographic Information File, mmCIF. It contains more information than previously used but now outdated PDB file format. The mmCIF file consists of category name and attribute name, their combination is called mmCIF token. Data are presented in two types, the first is key-value, in which the token is followed directly by a single value. The second type is tabular which is used when multiple values correspond to a single token (Burley, Berman et al. 2018).

Biological background: REP and RAYT

REP - repetitive extragenic palindromic sequences are non-coding bacterial transposable elements. They are found in high abundance in genomes and are very often clustered in so-called BIMEs - bacterial interspersed mosaic elements. RAYT - REP associated tyrosine transposases are usually flanked by two REPs and thus create transposable elements. Figure 18 depicts the genomic relationship between REP and RAYT (Nunvar, Huckova et al. 2010).

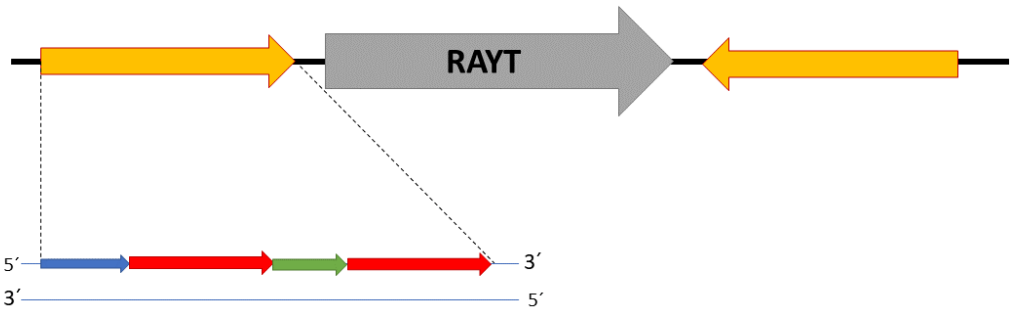

Figure 18: Schematic representation of REP-RAYT sequential relationship (Nunvar, Huckova et al. 2010).

RAYTs are related to the IS200/IS605 transposase family but show some distinct features. One of them is their inability to perform cleavage on double-stranded DNA, RAYT is active only on ssDNA. So far the only solved structure of *Escherichia Coli* RAYT with bound REP shows that the palindromic part is folded into a hairpin with an overhang in from of GTAG (Figure 19). Their nuclease activity has been studied extensively but their transposition function has not been proved

so far, it is only predicted based on their similarity with other transposable elements (Nunvar, Huckova et al. 2010).

Figure 19: To this day, the only solved X-ray structure (PDBid 4er8) of the bacterial REP-RAYT complex from Escherichia coli. RAYT (green) is bound to its cognate REP (orange) via 5'-GTAG recognition tetranucleotide (dark red), TT (blue) which causes the imperfection of the REP palindromicity are highlighted. Figure was created in ChimeraX (version 0.92) (Pettersen, Goddard et al. 2021).

Examined REP sequences are about 20 nucleotide long and each of them has characteristic GTAG recognition tetranucleotide on its 5' end. Biophysical studies have demonstrated that such sequences could potentially form multiple species with different topologies and they can coexist under physiological conditions in non-specified equilibrium, implying the possibility of regulation via those species (Charnavets, Nunvar et al. 2015). Various type of REPs are hypothesized to be involved in processes beyond transposition such as gene evolution, expression, mobility, transcription termination and supercoiling. Evolutionary studies point out that REPs are an old component of the bacterial genome (Di Nocera, De Gregorio et al. 2013).

Objectives of this thesis

The goal is to provide an introduction to structural features of nucleic acids and their formal description, specifically:

1. Review the architectures of nucleic acids, mainly DNA
2. Present the most widely used experimental techniques for study of nucleic acid structures
3. Overview structures of oligonucleotides with sequences related to the Repetitive Extragenic Palindromes, REPs
4. Experimentally characterize selected REP sequences with
    a. X-ray crystallography
    b. CD spectroscopy
5. Analyze selected DNA structures from the PDB using the novel nucleotide conformer classes NtC

Materials

Instruments

Instruments are listed in the Table 3 below.

Table 3: List of used instruments and their manufacturers.

Instrument	Manufacturer
Minicentrifuge VWR Galalxy Ministar C1413 (6000 RPM)	Thermo Scientific, USA
Centrifuge Microfuge 20	Beckman Coulter, USA
Thermoblock Biostep CHB-202	Thermo Scientific, USA
Laboratory Scale Ohaus Pioneer PA2102C	Thermo Scientific, USA
pH-meter OrionStar A211	Thermo Scientific, USA
Magnetic Steerer IKA RCTB S000	IKA, Germany
Pipettes Thermo Scientific F2	Thermo Scientific, USA
Crystallization robot Crystal Gryphon LCP	ARI, USA
Crystallization hotel Formulatrix RI1000	Formulatrix, USA
Chirascan Plus CD spectrometer	Applied Photophysics, UK
Bruker D8 Venture	Bruker, Germany
Synchrotron Bessy II	Helmholtz Zentrum Berlin, Germany

Chemicals

Oligonucleotides (Table 4) were synthesized, purified and purchased from company Sigma Aldrich. Oligonucleotides arrived in dry form and were kept in the fridge at 4 C. After they were diluted to final concentration the stock solutions were kept in the freezer at -20 ºC. Oligonucleotides for crystallization and CD experiments were only desalted as a purification step.

Table 4: Name of oligomers used in this study and their sequences

Name	Sequence (5' → 3')
Chom-18	GGTGGGGCTTGCCCCACC
Chom-18Br	GGTGGGGC(BrU)TGCCCCACC
Hpar-18	GGTGGGTCTTGACCCACC
Chom18mer_AT	GGTGGGGCATGCCCCACC
Chom18mer_TA	GGTGGGGCTAGCCCCACC
Chom18mer_CG	GGTGGGGCCGGCCCCACC
Chom18mer_GC	GGTGGGGCGCGCCCCACC
Chom18mer_TC	GGTGGGGCTCGCCCCACC

Chemicals for all experiments were purchased from Sigma Aldrich. Chemicals for optimization were at least > 98 % pure. Screening kits Natrix (Hampton Research, USA) and Nucleix (Qiagen, Germany) were used. Chemicals for buffer solutions were classified as *pro analysis* purity.

For the circular dichroism measuremetns the oligonucleotides were dissolved in phosphate buffer solution made by combining appropriate amount of two parts until pH of 7,4 is obtained:

Part I: 59.8 mM NaCl, 20 mM, Na_2HPO_4 and 0.1 mM Na_2EDTA

Part II: 79.8 mM NaCl, 20 mM, NaH_2PO_4 and 0.1 mM Na_2EDTA.

Methods

X-ray diffraction experiments

Oligonucleotide solution was prepared by diluting the lyophilized samples in distilled water to final concentration of 1 or 1.5 mM. Prior to experiments stock solutions were thawed at room temperature followed by heating up in thermoblock to 95 ºC for 10 minutes. After 10 minutes they were slowly cooled at laboratory temperature.

First approach was screening of the oligonucleotides using crystallization robot Gryphon (Art Robbins, USA), crystallization hotel Formulatrix RI1000 (Formulatrix, USA) and commercially available screening kits.

Crystallization robot is able to pipet in 96 well-plate, in each well there are three positions for one drop. Therefore, in each of the 96 unique conditions we can monitor three drops, they differ in ratio DNA stock/condition solution. Three ratios are 2:1, 1:1 and 1:2 to the final volume of 0,3 μL and volume of the well is 100 μL. 96 well-plate was in sitting drop setup. After the robot finished pipetting the sample, the plate was sealed to prevent drying up. Sealed plate was then inserted in a

crystallization hotel, where images of the drops were taken in defined time intervals. Temperature inside of the crystallization hotel was set to 20 ºC.

Hits in the screens held in the crystallization hotel were further optimized in a 24-well plate in the hanging drop setup. Volume of the drop was 3 µL and volume of the well was 1000 µL. Plates were stored in an incubator set to 20 ºC. Optimized crystals were fished out and freezed in liquid nitrogen. Sequences Chom18mer_AT, TA, CG, GC, TC did not require cryoprotection, the conditions already contain MPD in sufficient amounts (~25 % v/v) so that it can act as cryoprotectant.

Diffraction data were collected at BESSY II on beamline BL14.2 managed by Helmholtz-Zentrum Berlin (Mueller, Förster et al. 2015). The phase problem was solved using anomalous data from variant Chom-18Br. Data were processed using *XDS (Kabsch 2010)*, phasing was done using *AutoSol (Liebschner, Afonine et al. 2019)*, manual rebuilding was necessary using *Coot (Emsley, Lohkamp et al. 2010)*. Refinement was done with *phenix.refine (Afonine, Grosse-Kunstleve et al. 2012)*.

CD spectroscopy measurements

Oligonucleotides were thawed at room temperature followed by heating up in thermoblock to 95 ºC for 10 minutes. Concentration of samples was in the range of 5 to 20 µM. Spectra were obtained with spectropolarimeter Chirascan Plus (Applied Photophysics) and were measured in the range from 205 to 340 nm, with 1 nm step.

Analysis of DNA structures using NtC

Besides structure annotation in the refinement process we have selected DNA containing crystal structures with resolution better than 3.0 Å, PDB release of 5th November 2019. We searched their respective mmCIF token ndb_struct_na_base_pair.hbond_type_28 for values other than ‘19’,’20’ or ‘?’. These values denote presence of canonical Watson-Crick base pairing or unknown pairing pattern. Structures that met our criteria were uploaded to DNATCO server and assigned corresponding NtCs.

Evaluation of the fit to the electron density map and closeness of the investigated step and closest dinucleotide in the NtC class defining group of dinucleotides was carried out. The real-space correlation coefficients (RSCC) to the electron density were calculated using phenix.real_space_correlation with defined steps and the geometry closeness was represented as root-mean square deviation (r.m.s.d.) between the closest member of the golden set and the investigated step. Two values were plotted to the final scattergram of the RSCC versus r.m.s.d. (Afonine, Grosse-Kunstleve et al. 2012).

Results

X-ray structures

Obtained X-ray structures of sequences Chom-18, Chom-18Br and Hpar-18 were solved using data at crystallographic resolution of 2.7, 2.6 and 2.9 Å, respectively. In Table 5, the PDB codes of three discussed structures together with their resolutions are listed. Phase problem was solved experimentally with anomalous data from the Chom-18Br variant. During refinement, structures have been regularly uploaded to DNATCO server in order to monitor closeness of the unassigned steps with the closest NtC class. Geometries of the step defining parameters with a low torsional confal were attempted to alter so they would be assigned to a proper NtC class (Kolenko, Svoboda et al. 2020).

Table 5: Solved X-ray structures with their PDBid and crystallographic resolution.

Name	PDBid	Resolution
Chom-18	6ROS	2.7 Å
Chom-18Br	6ROR	2.6 Å
Hpar-18	6ROU	2.9 Å

Resolved crystal structures showed that all three variants form isomorphic antiparallel helix with consecutive T-T, BrU-T in the case of Chom-18Br, mismatches in the central region. Single DNA strand forms the asymmetric unit, the biological unit is generated by two-fold symmetry resulting in the duplex form. Full NtC assignment of the structures, Table 6, reveals the overall A-from character of the duplexes (Kolenko, Svoboda et al. 2020).

Table 6: NtC assignment of the reported X-ray structures.

6ROS dinucleotides		6ROR dinucleotides		6ROU dinucleotides	
DG_1_DG_2	AA08	DG_1_DG_2	AA08	DG_1_DG_2	AA04
DG_2_DT_3	AA00	DG_2_DT_3	AA00	DG_2_DT_3	AA00
DT_3_DG_4	AA08	DT_3_DG_4	AA00	DT_3_DG_4	AA00
DG_4_DG_5	AA04	DG_4_DG_5	AA04	DG_4_DG_5	NANT
DG_5_DG_6	AA00	DG_5_DG_6	AA00	DG_5_DG_6	AA08
DG_6_DG_7	AA10	DG_6_DG_7	AA01	DG_6_DT_7	AA11
DG_7_DC_8	AA08	DG_7_DC_8	AA08	DT_7_DC_8	AA08
DC_8_DT_9	AA00	DC_8_BRU_9	AA08	DC_8_DT_9	AA00
DT_9_DT_10	AA08	BRU_9_DT_10	AA08	DT_9_DT_10	AA08
DT_10_DG_11	NANT	DT_10_DG_11	NANT	DT_10_DG_11	NANT
DG_11_DC_12	NANT	DG_11_DC_12	NANT	DG_11_DA_12	NANT
DC_12_DC_13	BA08	DC_12_DC_13	BA08	DA_12_DC_13	NANT
DC_13_DC_14	AA00	DC_13_DC_14	AA00	DC_13_DC_14	AA00
DC_14_DC_15	AA08	DC_14_DC_15	AA08	DC_14_DC_15	AA08
DC_15_DA_16	AA06	DC_15_DA_16	AA06	DC_15_DA_16	AA06
DA_16_DC_17	AA08	DA_16_DC_17	AA08	DA_16_DC_17	AA08
DC_17_DC_18	AB05	DC_17_DC_18	AB05	DC_17_DC_18	AA00

Although crystals of other variants of Chom18 (Chom18mer_AT, TA, GC, CG and TC, Table 4) have been successfully optimized, because of poor quality of in-house diffraction images caused by issues with the cryo-pump on the diffractometer Bruker D8 Venture at the Center of Molecular Structure at the Institute of Biotechnology CAS, they have been freezed in liquid nitrogen and stored for future data collection. High concentration of MPD (~ 25 % v/v) is cryoprotective, therefore no further cryoprotection was necessary.

CD spectroscopy

Circular dichroism spectra of the Chom-18 and Hpar-18 crystal structure sequences were measured, Figure 20 and 21 show CD curves measured in different solutions. Melting CD spectra for Chom-18 and Hpar-18 are shown in Figure 22 and 23 (Kolenko, Svoboda et al. 2020). Data for Hpar-22 and Chom-22 (same sequence as the 18-mers but preceded by GTAG on 5'- end) have already been published elsewhere (Charnavets, Nunvar et al. 2015).

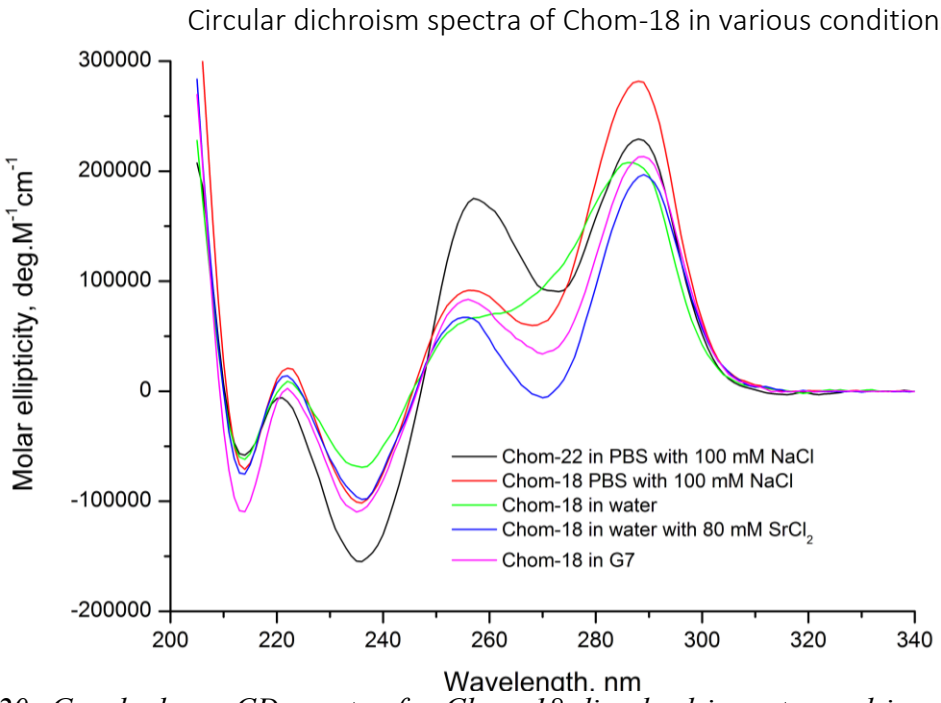

Figure 20: Graph shows CD spectra for Chom-18 dissolved in water and in other buffers and solutions. G7 is crystallization condition in Natrix screen kit (Kolenko, Svoboda et al. 2020). For comparison, the CD spectra for Chom-22 are shown (Charnavets, Nunvar et al. 2015).

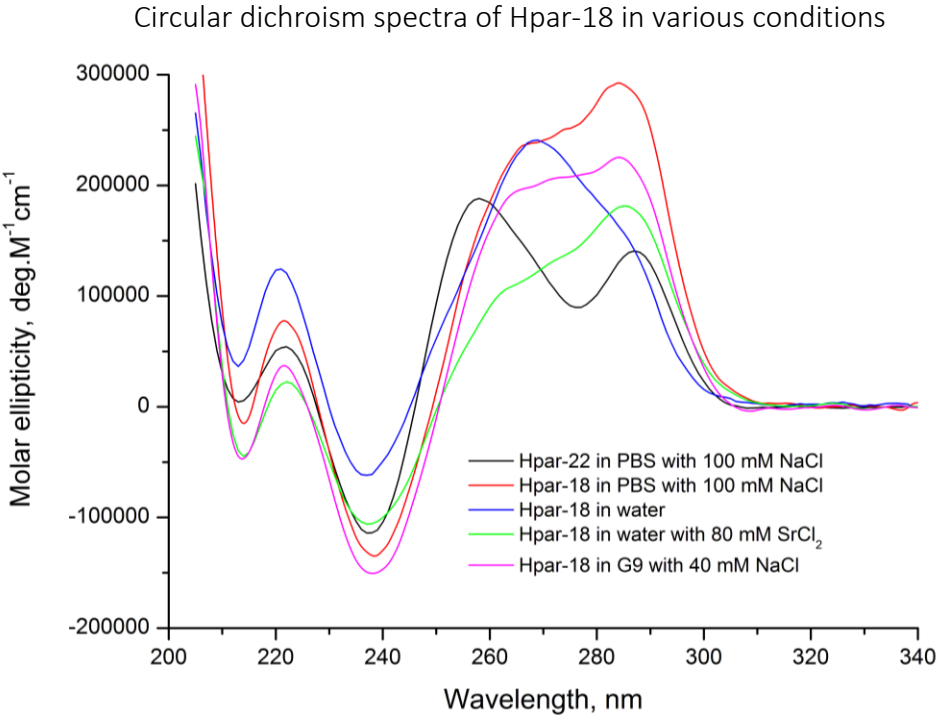

Figure 21: CD spectra for Hpar-18 in various conditions, similar to previous Chom-18 figure (Kolenko, Svoboda et al. 2020). G9 is the crystallization condition (Natrix) and data for Hpar-22 are shown (Charnavets, Nunvar et al. 2015).

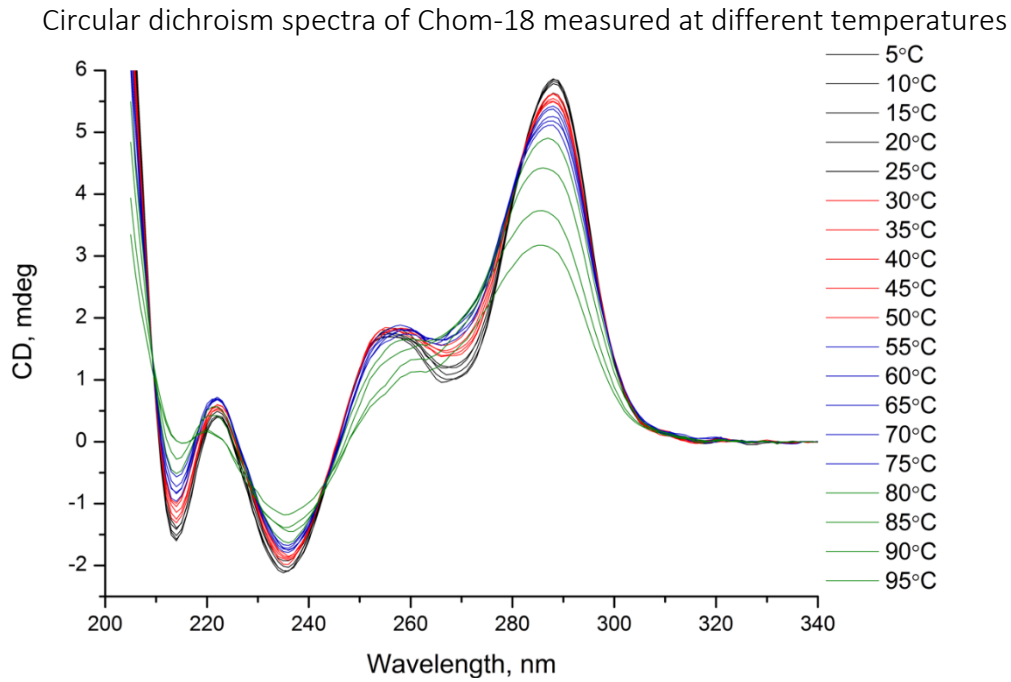

Figure 22: CD spectra of samples Chom-18 registered at different temperatures (Kolenko, Svoboda et al. 2020).

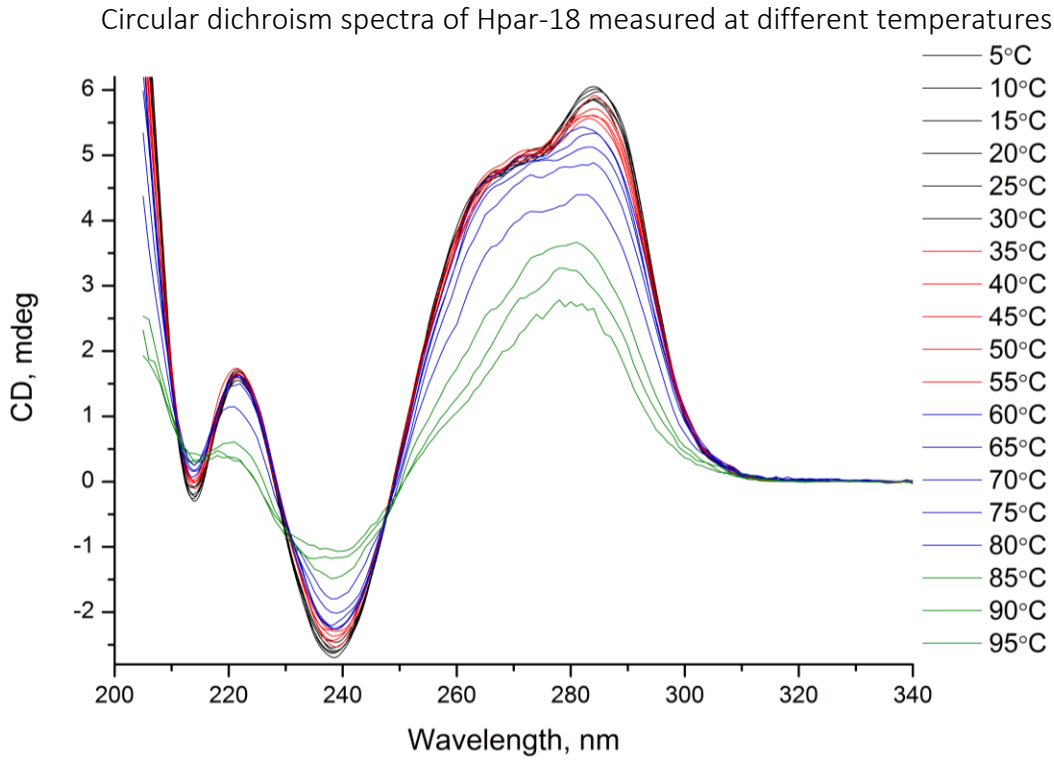

Figure 23: CD spectra of Hpar-18 at different temperatures (Kolenko, Svoboda et al. 2020).

Effect of titration on the CD spectra of Chom-18 and Hapr-18 with Sr^{2+} is shown in Figure 24 and 25.

Titration of Chom-18 and respective change in the CD spectrum

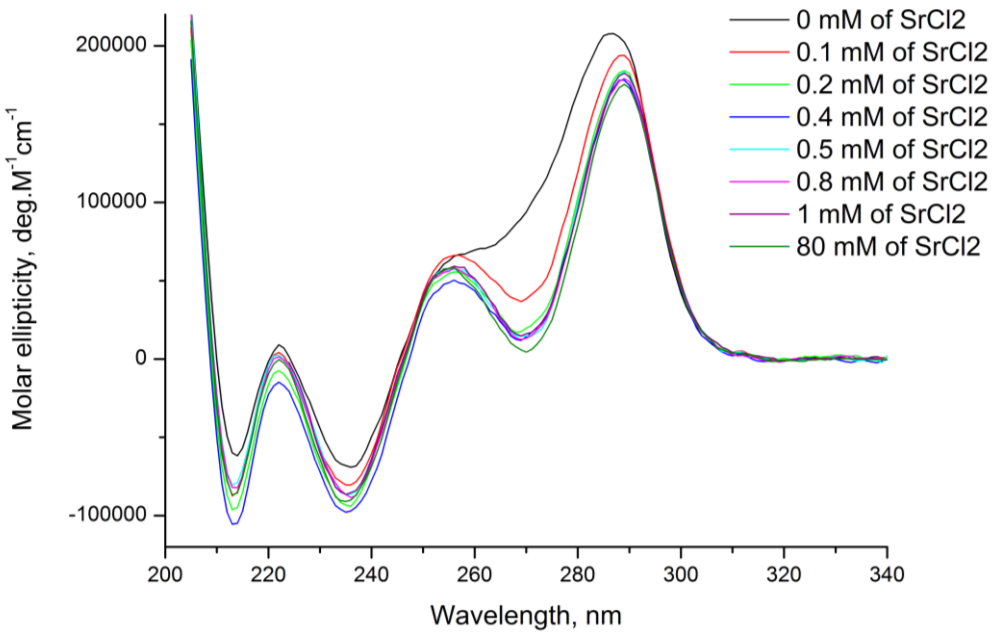

Figure 24: Change of the CD spectrum of Chom-18 during titration with Sr^{2+} (Kolenko, Svoboda et al. 2020).

Titration of Hpar-18 and respective change in the CD spectrum

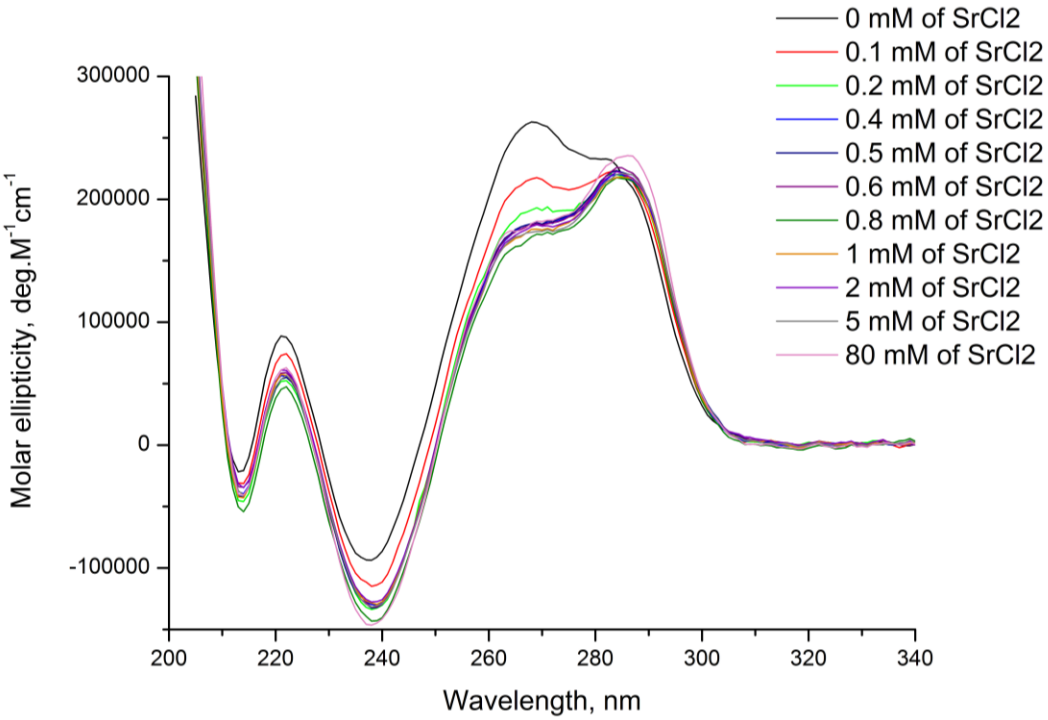

Figure 25: Change of the CD spectrum of Hpar-18 during titration with Sr^{2+} (Kolenko, Svoboda et al. 2020).

Analysis of the non-canonically paired dinucleotides

Presence of two consecutive T-T mismatched base pairs in the solved structures motivated us to perform more extensive structural analysis of geometries of dinucleotides involved in the non-canonical base pairing. We have retrieved 1094 paired dinucleotides, in which at least one pair is classified as non-canonical (mismatched) according to a value of the token ndb_struct_na_base_pair.hbond_type_28 in the mmCIF files. This dataset includes antiparallel duplexes, parallel duplexes, and tetraplexes of DNA and their complexes with proteins. The incidences of non-canonical base pairs are listed in Table 7.

Table 7: Incidences of non-canonical base pairs found in selected DNA containing structures in PDB database. A-T and C-G base pairs in this table are the reverse Watson-Crick.

Base pair	A-A	A-C	A-G	A-T	C-C	C-G	C-T	G-G	G-T	T-T
Antiparallel	16	8	175	193	0	127	14	72	141	42
Parallel	34	0	0	1	115	0	0	153	0	3

T-T mismatches were found mainly in the antiparallel duplexes, 42 cases, and 3 cases in the parallel ones. G-G and even more strikingly C-C base pairs were found almost exclusively in the parallel strands, they are often in structures of G-tetraplexes and i-motifs. Multiplexed structural elements topologically allow incorporation of parallel strands with less effort than duplexes.

Considered one of the most stable and double helix least disrupting non-canonical base pair, G-T (Pan, Sun et al. 2006) was observed only in the antiparallel structures, 141 cases. Similarly, A-G, A-C, A-T, C-G and C-T were found almost in all cases, apart from one, in the antiparallel orientation and not in the parallel ones. A-A pair was found twice as much in the parallel (34 cases) than in the antiparallel (16 cases).

In order to gain insight into conformation of the dinucleotide steps involved in the non-canonical base pair we utilized flexibility and robustness of NtC classes. The assigned NtC corresponding to the four steps flanking non-canonical base pair from both ends have been accounted for in the further analysis. The analyzed fragment is depicted in Figure 26.

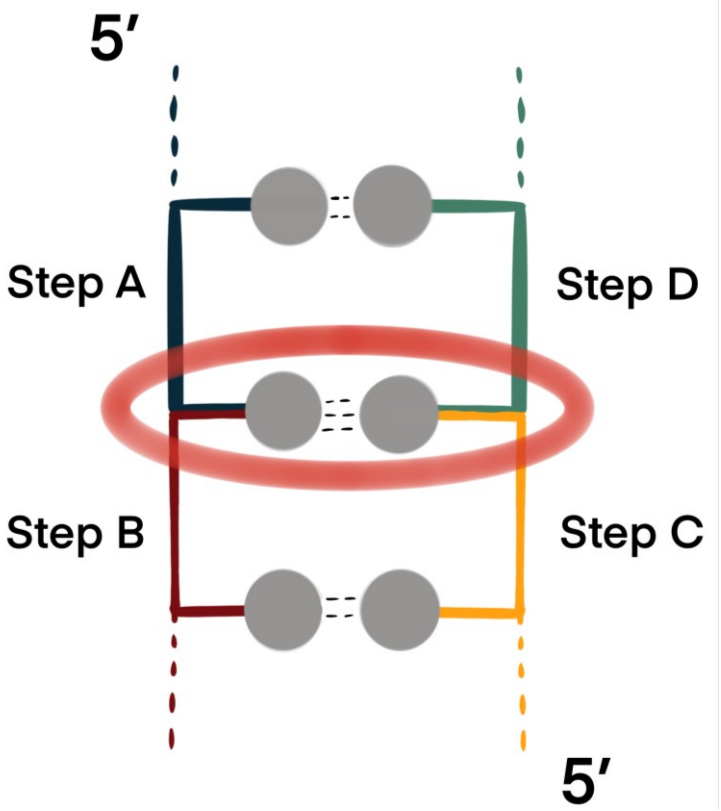

Figure 26: Schematic representation of the analyzed DNA fragment. The non-canonical base pair is marked red, the four surrounding dinucleotide steps as Step A to D. The direction of both strands are indicated by labelled 5' ends. Obviously, in the case of parallel strands, direction of one strand is reversed.

Distribution of NtCs around non-canonically paired bases is illustrated in Table 8. The most common steps were unassigned NANT (38.0 %), followed by BB00 (21.2 %) and BBS1 (6.0 %). Minor B-form NtCs were assigned in relatively small quantities. NtC classes representing A-form of duplex were found only in around 5 % of cases, mostly in mismatches with thymine and in C-C base pairs. For DNA relatively rare classes representing open or intercalated conformations such as OP15 and IC06 were found scarcely. Same applies for AB01, AB03, AB05, BA01 and BA05 etc.

Scattergrams for the most populated NtC classes in mismatched base pair dataset compared with distribution in all dinucleotide steps found in the PDB (all resolutions, (Cerny, Bozikova et al. 2020)) are shown in Figure 27. The scattergrams with the data from the entire PDB are shown in contour representation rather than as individual points.

Table 8: Distribution of the NtC classes in the mismatched pairs. For clarity, only NtC classes with total incidences higher or equal to 5 are shown. A-T and C-G are the Hoogsteen-paired and reverse Watson-Crick base pairs.

	Mismatched base pair									
	A-A	A-C	A-G	A-T	C-C	C-G	C-T	G-G	G-T	T-T
AA00	1	0	3	12	0	15	1	22	36	30
AA02	0	0	4	8	0	2	2	2	23	0
AA04	0	0	0	0	0	1	0	0	10	0
AA08	0	0	6	1	0	1	0	2	7	3
AA09	6	0	0	1	0	4	0	2	2	0
AA01	0	0	0	0	0	5	0	0	2	3
AA07	1	0	0	0	6	0	0	1	0	0
AA12	0	0	0	0	20	0	0	0	0	0
AB01	5	1	8	18	0	24	4	36	31	11
AB02	0	0	3	0	0	2	0	0	1	1
AB03	1	0	5	2	0	19	3	2	8	2
AB05	10	0	4	14	3	2	0	17	5	6
BA01	2	0	2	1	6	21	2	6	28	4
BA05	3	7	28	11	0	10	3	4	33	11
BA08	0	1	11	1	0	5	1	8	3	0
BA10	0	0	2	0	0	4	0	0	1	3
BA13	0	0	0	4	0	1	0	4	0	1
BB00	34	2	233	101	43	96	14	248	118	29
BB01	7	0	19	18	0	27	4	11	41	12
BB17	0	0	2	1	0	2	0	8	0	0
BB02	7	0	4	13	0	10	0	5	6	3
BB03	1	0	2	1	0	1	0	2	0	0
BB16	13	0	33	8	19	9	0	20	1	3
BB04	2	1	27	43	0	22	2	8	24	6
BB07	0	0	27	34	0	15	0	2	18	3
BB10	0	1	2	2	0	10	0	2	1	1
BB12	3	0	1	1	0	5	0	0	4	0
BB13	0	0	4	1	0	4	0	0	0	0
BB15	2	0	0	0	0	7	2	5	9	1
IC06	0	0	4	0	14	3	0	6	0	0
OP15	19	0	16	0	6	0	0	41	0	2
BB1S	0	0	0	0	0	0	0	16	0	0
BB2S	0	0	0	5	0	0	0	10	0	0
BBS1	9	4	19	118	1	0	0	108	0	1
ZZ1S	0	0	1	0	0	1	0	0	18	0
ZZ2S	0	0	1	0	0	1	0	0	8	0
NANT	72	14	227	350	342	166	18	296	116	44

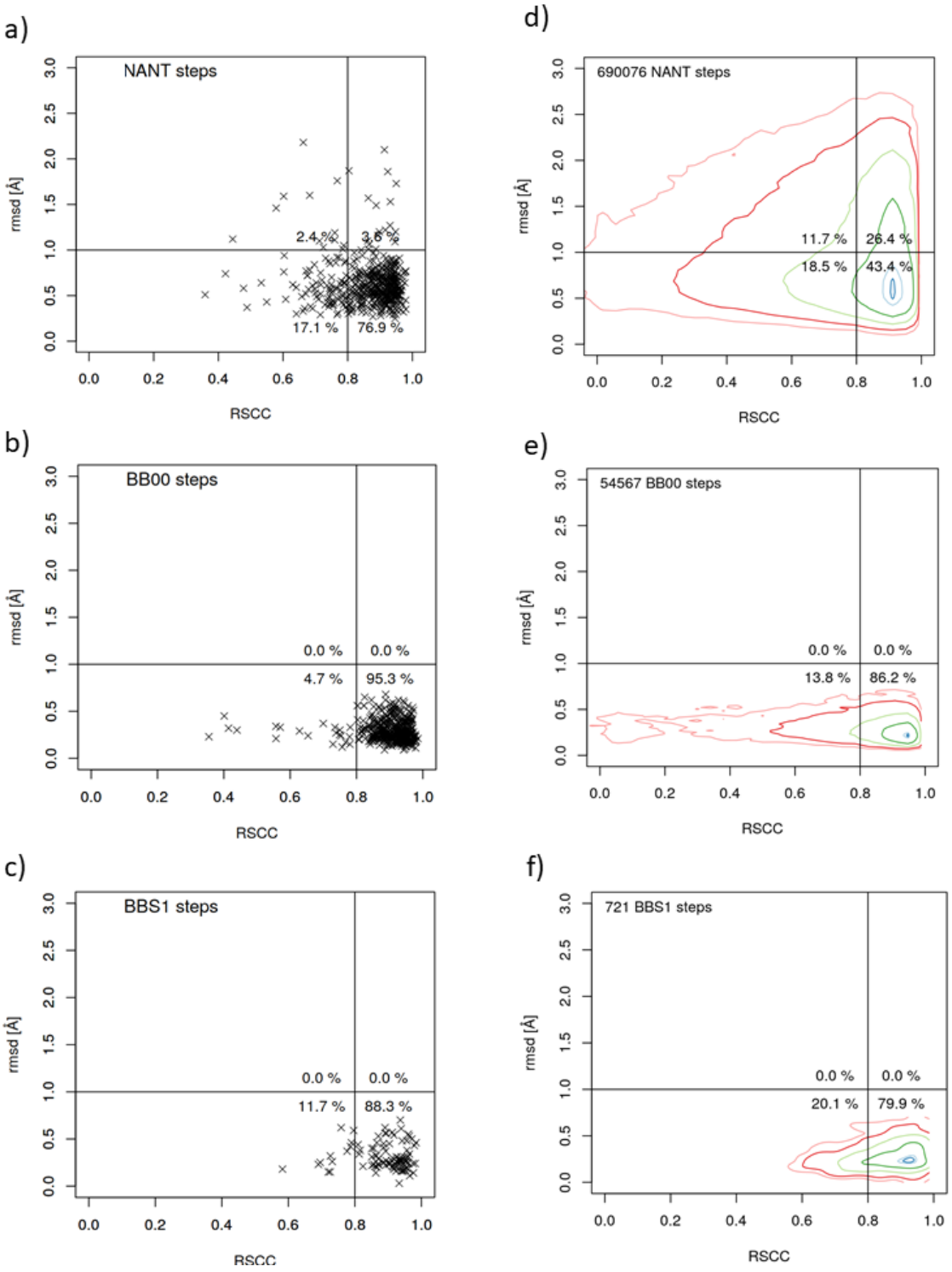

Figure 27: Scattergrams showing the fit to the electron density (RSCC) and r.m.s.d. for three most populated NtC classes (NANT, BB00 and BBS1) in the mismatched base pairs (a-c) and for the dinucleotide steps in the entire PDB database (d-f) (Cerny, Bozikova et al. 2020).

Discussion

X-ray structures

X-ray diffraction experiment revealed that sequences Chom-18 and Hpar-18 crystallized into a duplex with two consecutive mismatched base pairs in the central region (Figure 28). Two consecutive T-T mismatches are the first of its kind reported in the PDB. Experimental phasing was necessary due to no available molecular model.

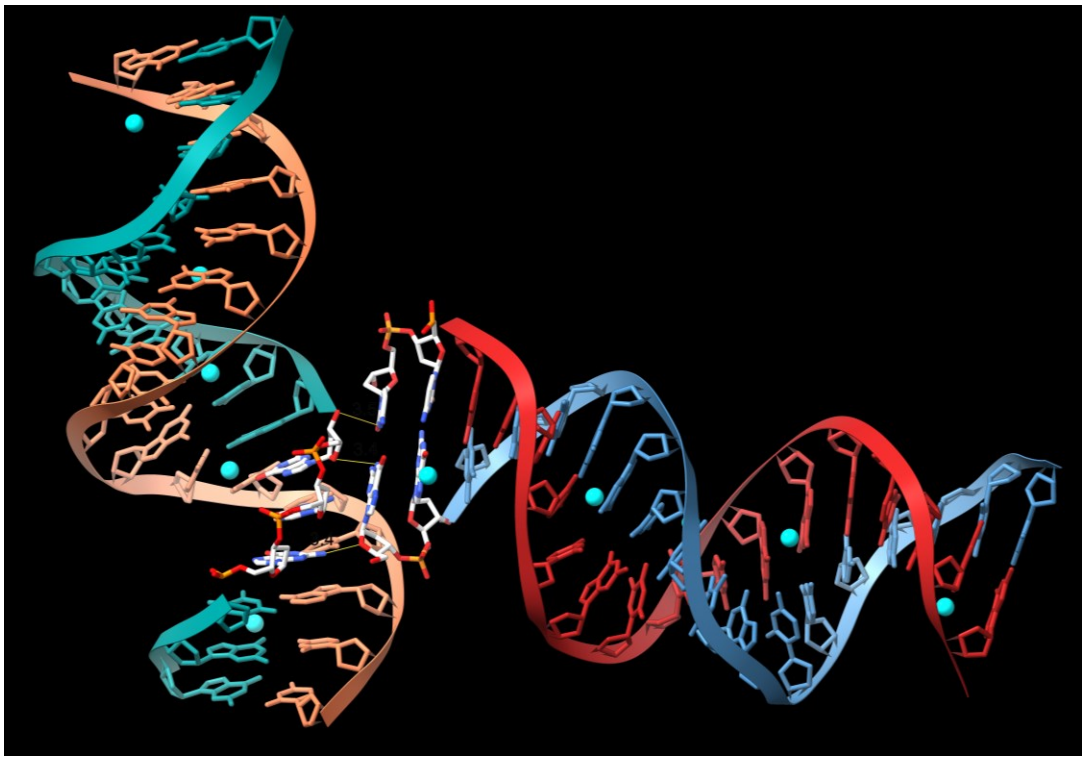

Figure 28: Crystal packing of two duplexes in the structure 6ror is shown. The cyan spheres represent Sr^{2+} cations. Distances between atoms of the two molecules are given in Å. Figure was created in ChimeraX (version 0.92) (Pettersen, Goddard et al. 2021).

Annotation with NtC showed the A-form structural features. Steps involved in the mismatched T-T (BrU-T) are assigned AA00, AA00 (AA08 in 6ROR) followed by NANT. According to statistics on the DNATCO server, AA08 is the second most populated A-form class after canonical AA00, therefore T-T (BrU-T) mismatched in three reported structures does not significantly deform the geometry of the duplex. Electron density around central TT region followed by nucleotides T10-C13 is considered to be quite poor, refinement of this fragment turned out to be difficult. However, it was substantially aided by the use of NtC assignment in critical torsions.

The main crystal packing interactions are depicted in Figure 28. Duplexes are packed via interactions of nucleotides G4, G5 and G6 of one strand and G1* and C18** of second symmetry related duplex, a packing mode that was observed in other DNA crystal structures. Atoms of two helices are around 3.5 Å apart in the area of contact.

CD spectroscopy

The G-rich stretches in all sequences suggest formation of G-tetraplexes and the measured CD spectra provide evidence for the existence of mixtures of helical species and anti-parallel G-tetraplexes. The co-existence of several molecular species in solution of the analyzed oligonucleotides was confirmed by the SVD analysis of temperature-dependent CD spectra in various buffer solutions. Together with the absence of the isodichroic point in the titration spectra we can assess that there are at least three to four conformational species formed in the solution of Chom-18 and Hpar-18.

CD spectra measured in different buffers (Figures 20 and 21) show similar behavior in the range of approximately 200-240 nm. Chom-18 curves share analogous characteristic peaks in solutions containing significant amounts of cations (Na^+, Sr^{2+} etc.), mainly negative ones around 240 and 270 nm and positive around 220, 260 ad 290 nm. Reported peaks are positioned similarly as in the case of Chom-22, their intensities are comparable, apart from the peak near 260 nm. However the spectrum of Chom-18 is considerably flattened when measured just in distilled water, no cations. Hpar-18 displays similar characteristic peaks in solution with cations as Chom-18. Although negative peak around 270 nm is not discernable, positive peaks around 220, 270 and 290 nm and negative ones around 240 nm are easily distinguishable. The Hpar-18 and Hpar-22 spectra measured in the same buffer (PBS with added NaCl) look different at the first sight but the peak positions are comparable.

The G-tetraplex signal was detected in the circular dichroism spectroscopic measurements. Particularly the signal for antiparallel G-tetraplex. This opens the possibility of presence of the higher order architectures in the solution. Possible topological arrangements of the solution form are depicted in Figure 29. Molecular dynamics simulation (not part of this thesis) hinted that conformations c)-f) in Figure 29 are unlikely present in solution for sufficient amount of time and/or in high enough concentration to successfully nucleate during early stage of crystallization. This is presumably due to composition of the loop, in each case it is composed of single nucleotide (T3). This seems to cause excessive steric strain on the G-tetraplex architecture. We acknowledge only bimolecular G-tetraplexes on account of mass spectroscopic measurements and capillary electrophoretic experiments (neither of them shown since they require further optimization) with longer variant Chom-22. They confirmed that only dimeric or lighter particles are present in the solution.

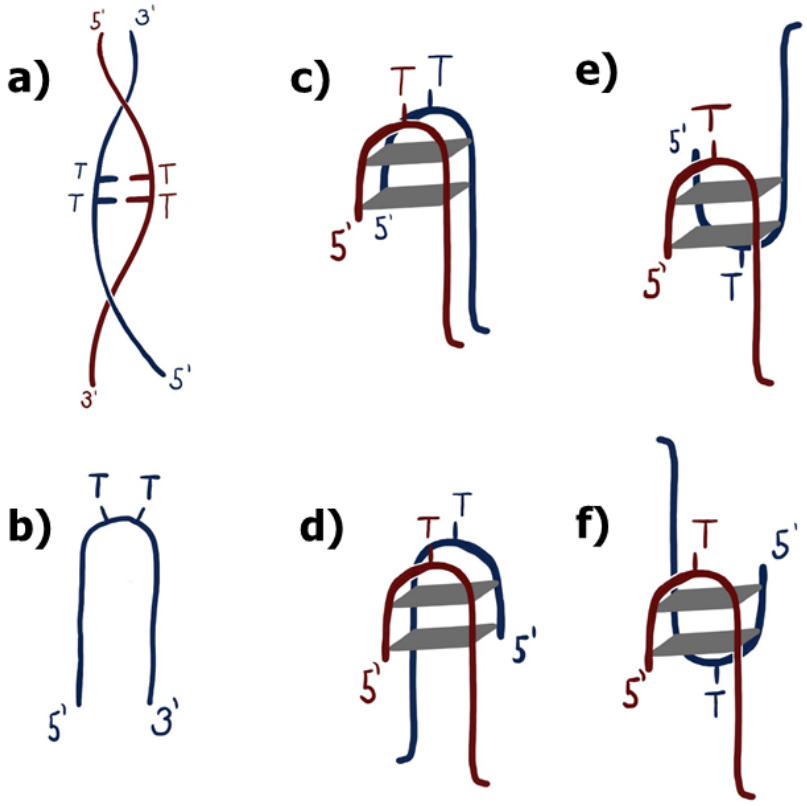

Figure 29: Hypothesized topologies of Chom18, Chom18-Br and Hpar18 in solution (Kolenko, Svoboda et al. 2020).

Melting CD curves in Figures 22 and 23 establish the fact that conformers contributing to them are stable up to 75-80 ºC. Melting temperatures this high is not a typical for duplex species (Figure 29 a) and b)), they rather indicate presence of a mixture of more thermodynamically stable G-tetraplexes (Figure 29 c-f).

The double helical form of all three crystallized oligonucleotide sequences is to some extent counter-intuitive. The duplexes contain supposedly destabilizing double T-T mismatches, they crystallized in the presence of Sr^{2+}, the cation that to the best of our knowledge induces formation of guanine tetraplexes, and the CD spectra are suggestive of tetraplex species in solution. Therefore, we have decided to investigate the role of Sr^{2+} on the conformational space of Chom-18 and Hpar-18 in a greater detail. There is a noticeable change in the CD spectra of both oligomers even after addition of the smallest amount of $SrCl_2$ that does not change significantly at higher concentration of Sr^{2+} ($SrCl_2$ up to 4 mM). These facts can be accredited to easily induced formation of G-tetraplexes, yet the process of crystallization of the oligonucleotides induced the double helical form indicating the complex nature of equilibria of biomolecules.

Analysis of structures containing mismatched base pairs

T-T mismatches, also present in our X-ray structures, were found in double helices either unassigned (NANT – 44 cases) or assigned dominantly AA00 (30 cases) and BB00 (29 cases) classes, followed by BB01, AB01 and AB05. Although T-T base pairs can form three distinct base pair types according to the Saenger pairing nomenclature (numbers 12, 13 and 16 in Figure 8), 42 out of 45 cases were denoted as number 16 and only 3 as number 12 and none as number 13.

Generally, the guanine containing mismatched base pairs (A-G, G-G and G-T) were found in the most cases. Confirming the fact that guanine can entertain non-canonical base pairs via non-Watson-Crick edges rather effortlessly compared to the other bases. G-G and even more strikingly C-C base pairs were found mostly in the parallel strands, they are often in structures of G-tetraplexes and i-motifs. Multiplexed structural elements topologically allow incorporation of parallel strands with less effort than duplexes. G-G pairs were, apart from NANT, assigned BB00 and BBS1 NtC classes. These two classes are known to be crucially involved in the building of G-tetraplexes.

Distribution of the NtC classes in the non-canonical base pairing is not consistent with the statistics found for all dinucleotides on the DNATCO website. Although incidences of the AAA NtC classes for dinucleotides involved in both mismatched and Watson-Crick pairs are similar, they differ quitter significantly for the BB00 or NANT NtCs. We found that BB00 class forms only 21.2 % of mismatched base pairs but full 33.6 % in the W-C pairs; the corresponding fractions for the NANT dinucleotides are 38.0 % and 18.9 %. As mentioned above, high counts of BBS1 can be partly attributed to G-tetraplex structures, which are known to contain such a class. Large set of structures containing non-canonical A-T base pair is surprisingly assigned to the BBS1 class.

However the high occurrences of C-G and A-T base pairs in non-canonical (non-Watson-Crick) arrangements caught our attention and therefore we have decided to manually check 50 randomly chosen structures that claimed to contain them. Most of them, while being classified as non-canonical, formed almost perfect Watson-Crick base pair.

Scattergrams on Figure 27 comprehensively visualizes geometrical differences between step of interest and the nearest step in the golden set and their fit to the electron density. In the case of assigned classes (BB00 or BBS1) there is evident correlation between values of RSCC and r.m.s.d. where most of the steps are located in the region of high electron density correlation and small geometrical difference r.m.s.d. Unassigned steps (NANT) displays larger dispersion of points in the scattergram. Region of high RSCC (0.8 - 1) and low r.m.s.d. (less than 1 Å) could indicate that 76.9 % of NANT mismatched steps could potentially be re-refined in the critical torsions with the knowledge of NtC. Scattergrams for steps assigned to all available structures in PDB show similar distribution for BB00 and BBS1 NtC classes but unassigned (NANT) mismatched steps display lower geometrical differences then all unassigned steps.

Conclusions

We have characterized selected DNA oligomers with sequences related to bacterial non-coding elements called Repetitive Extragenic Palindromes, REPs by means of crystallography and spectroscopy. We obtained crystal structures of three sequences, Chom-18, Chom18-Br and Hpar-18. They crystallized into isomorphic double helical antiparallel duplexes (Figure 28). The refinement process was helped by the knowledge of the NtC classes assigned to incompletely refined dinucleotides. The sequential similarity of the analyzed Chom-18 and Hpar-18 oligonucleotides has been projected in their respective solution CD spectra as well as to their crystal structures (Figures 20 - 25). They both occupy complex conformational space in solution ranging from unimolecular hairpins to duplexes and bimolecular G-tetraplexes. G-tetraplex character of the CD spectra could be ascribed to a relatively high extinction coefficient of G-tetraplexes compared to that of duplexes or hairpins. Practically, it means that in solution, the actual ratio of conformers could be overshadowed by higher extinction coefficients of some structural species occurring at small fractions.

Two consecutive T-T mismatch in the central region of the crystal structures led us to the subsequent analysis of selected mismatched base pairs containing crystal structures. It revealed that the most common non-canonical base pairs are A-T, A-G and G-G (Table 8). Close to a half of mismatched base pair containing steps are not assigned to any NtC class, one fifth is assigned to the BB00 class. Discrepancies in the mmCIF files impose obstacles for a proper base pairing analysis, there is a need for a systematic revision. We conclude that non-canonical base pair cause no observable deformation in the step defining parameters. Reported scattergram (Figure 27-a)) for the class NANT of unassigned dinucleotides hints suggests that geometries of a significant part (more than 3/4) of the analyzed dinucleotide steps could be made compliant with the known conformations by a proper re-refinement process.

References

Abou Assi, H., M. Garavís, C. González and M. J. Damha (2018). "i-Motif DNA: structural features and significance to cell biology." Nucleic Acids Res **46**(16): 8038-8056.

Adrian, M., B. Heddi and A. T. Phan (2012). "NMR spectroscopy of G-quadruplexes." Methods **57**(1): 11-24.

Afonine, P. V., R. W. Grosse-Kunstleve, N. Echols, J. J. Headd, N. W. Moriarty, M. Mustyakimov, T. C. Terwilliger, A. Urzhumtsev, P. H. Zwart and P. D. Adams (2012). "Towards automated crystallographic structure refinement with phenix.refine." Acta Crystallogr D Biol Crystallogr **68**(Pt 4): 352-367.

Al-Hashimi, H. M. (2013). "NMR studies of nucleic acid dynamics." J Magn Reson **237**: 191-204.

Avery, O. T., C. M. MacLeod and M. McCarthy (1944). "Studies on the chemical nature of the substance inducing transformation of pneumococcal types: induction of transformation by a desoxyribonucleic acid fraction isolated from Pneumococcus Type III." J. Exp. Med. **79**: 137-158.

Berman, H. M., J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov and P. E. Bourne (2000). "The Protein Data Bank." Nucleic Acids Res. **28**(1): 235-242.

Biedermannova, L. and B. Schneider (2016). "Hydration of proteins and nucleic acids: Advances in experiment and theory. A review." Biochim Biophys Acta **1860**(9): 1821-1835.

Blackburn, G. M. and M. J. Gait (2006). Nucleic Acids in Chemistry and Biology. Oxford, IRL Press at Oxford Univ. Press.

Blanchet, C. E. and D. I. Svergun (2013). "Small-Angle X-Ray Scattering on Biological Macromolecules and Nanocomposites in Solution." Annual Review of Physical Chemistry **64**(1): 37-54.

Brzezinski, K., A. Brzuszkiewicz, M. Dauter, M. Kubicki, M. Jaskolski and Z. Dauter (2011). "High regularity of Z-DNA revealed by ultra high-resolution crystal structure at 0.55 A." Nucleic Acids Res **39**(14): 6238-6248.

Burge, S., G. N. Parkinson, P. Hazel, A. K. Todd and S. Neidle (2006). "Quadruplex DNA: sequence, topology and structure." Nucleic Acids Res **34**(19): 5402-5415.

Burley, S. K., H. M. Berman, C. Bhikadiya, C. Bi, L. Chen, L. Di Costanzo, C. Christie, K. Dalenberg, J. M. Duarte, S. Dutta, Z. Feng, S. Ghosh, D. S. Goodsell, R. K. Green, V. Guranović, D. Guzenko, B. P. Hudson, T. Kalro, Y. Liang, R. Lowe, H. Namkoong, E. Peisach, I. Periskova, A. Prlić, C. Randle, A. Rose, P. Rose, R. Sala, M. Sekharan, C. Shao, L. Tan, Y.-P. Tao, Y. Valasatava, M. Voigt, J. Westbrook, J. Woo, H. Yang, J. Young, M. Zhuravleva and C. Zardecki (2018). "RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy." Nucleic Acids Research **47**(D1): D464-D474.

Černý, J., P. Božíková, M. Malý, M. Tykac, L. Biedermannova and B. Schneider (2020). "Structural alphabets for conformational analysis of nucleic acids available at dnatco.datmos.org." Acta Crystallographica Section D Structural Biology **76**: 805-813.

Cerny, J., P. Bozikova, J. Svoboda and B. Schneider (2020). "A unified dinucleotide alphabet describing both RNA and DNA structures." Nucleic Acids Res.

Charnavets, T., J. Nunvar, I. Necasova, J. Volker, K. J. Breslauer and B. Schneider (2015). "Conformational diversity of single-stranded DNA from bacterial repetitive extragenic palindromes: Implications for the DNA recognition elements of transposases." Biopolymers **103**(10): 585-596.

Cox, R. A. (1970). "Conformation of nucleic acids and the analysis of the hypochromic effect." Biochem J **120**(3): 539-547.

Crick, F. H. C. and J. D. Watson (1954). "The Complementary Structure of Deoxyribonucleic Acid." Proc. Roy. Soc. (London) Ser. A **223**: 80-96.

Cuenoud, B. and J. W. Szostak (1995). "A DNA metalloenzyme with DNA ligase activity." Nature **375**(6532): 611-614.

Di Nocera, P. P., E. De Gregorio and F. Rocco (2013). "GTAG- and CGTC-tagged palindromic DNA repeats in prokaryotes." BMC Genomics **14**: 522.

Dohm, J. A., M. H. Hsu, J. R. Hwu, R. C. Huang, E. N. Moudrianakis, E. E. Lattman and A. G. Gittis (2005). "Influence of ions, hydration, and the transcriptional inhibitor P4N on the conformations of the Sp1 binding site." J Mol Biol **349**(4): 731-744.

Drew, H. R., R. M. Wing, T. Takano, C. Broka, S. Tanaka, K. Itakura and R. E. Dickerson (1981). "Structure of a B-DNA dodecamer: conformation and dynamics." Proc.Natl.Acad.Sci.USA **78**(4): 2179-2183.

Ducruix, A. and R. Giege, Eds. (1992). Crystallization of Nucleic Acids and Proteins -- A Practical Approach. Oxford, IRL Press at Oxford University Press.

Dyer, K. N., M. Hammel, R. P. Rambo, S. E. Tsutakawa, I. Rodic, S. Classen, J. A. Tainer and G. L. Hura (2014). "High-throughput SAXS for the characterization of biomolecules in solution: a practical approach." Methods in molecular biology (Clifton, N.J.) **1091**: 245-258.

Emsley, P., B. Lohkamp, W. G. Scott and K. Cowtan (2010). "Features and development of Coot." Acta Crystallogr D Biol Crystallogr **66**(Pt 4): 486-501.

Fallmann, J., S. Will, J. Engelhardt, B. Grüning, R. Backofen and P. F. Stadler (2017). "Recent advances in RNA folding." Journal of Biotechnology **261**: 97-104.

Frank-Kamenetskii, M. D. and S. M. Mirkin (1995). "Triplex DNA structures." Annu Rev Biochem **64**: 65-95.

Franklin, R. E. and R. G. Gosling (1953). "Molecular configuration in sodium thymonucleate." Nature **171**: 740-741.

Ghosh, M., N. V. Kumar, U. Varshney and K. V. Chary (1999). "Structural characterisation of a uracil containing hairpin DNA by NMR and molecular dynamics." Nucleic Acids Res **27**(19): 3938-3944.

Hershey, A. and M. Chase (1952). Cold Spring Harbor Symp. Quant. Biol. **16**: 445-456.

Jeddi, I. and L. Saiz (2017). "Three-dimensional modeling of single stranded DNA hairpins for aptamer-based biosensors." Scientific Reports **7**(1): 1178.

Kabsch, W. (2010). "Xds." Acta Crystallographica Section D: Biological Crystallography **66**(2): 125-132.

Kolenko, P., J. Svoboda, J. Cerny, T. Charnavets and B. Schneider (2020). "Structural variability of CG-rich DNA 18-mers accommodating double T-T mismatches." Acta Crystallographica Section D Structural Biology **76**: 1233-1243.

Kypr, J., I. Kejnovska, D. Renciuk and M. Vorlickova (2009). "Circular dichroism and conformational polymorphism of DNA." Nucleic Acids Res **37**(6): 1713-1725.

Lawson, D. M., P. J. Artymiuk, S. J. Yewdall, J. M. Smith, J. C. Livingstone, A. Treffry, A. Luzzago, S. Levi, P. Arosio, G. Cesareni and et al. (1991). "Solving the structure of human H ferritin by genetically engineering intermolecular crystal contacts." Nature **349**(6309): 541-544.

Leontis, N. B. and E. Westhof (2001). "Geometric nomenclature and classification of RNA base pairs." RNA **7**(4): 499-512.

Liang, X., H. Kuhn and M. D. Frank-Kamenetskii (2006). "Monitoring Single-Stranded DNA Secondary Structure Formation by Determining the Topological State of DNA Catenanes." Biophysical Journal **90**(8): 2877-2889.

Liebschner, D., P. V. Afonine, M. L. Baker, G. Bunkóczi, V. B. Chen, T. I. Croll, B. Hintze, L. W. Hung, S. Jain, A. J. McCoy, N. W. Moriarty, R. D. Oeffner, B. K. Poon, M. G. Prisant, R. J. Read, J. S. Richardson, D. C. Richardson, M. D. Sammito, O. V. Sobolev, D. H. Stockwell, T. C. Terwilliger, A. G. Urzhumtsev, L. L. Videau, C. J. Williams and P. D. Adams (2019). "Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix." Acta Crystallogr D Struct Biol **75**(Pt 10): 861-877.

Mooers, B. (2008). "Crystallographic studies of DNA and RNA." Methods (San Diego, Calif.) **47**: 168-176.

Mueller, U., R. Förster, M. Hellmig, F. U. Huschmann, A. Kastner, P. Malecki, S. Pühringer, M. Röwer, K. Sparta, M. Steffien, M. Ühlein, P. Wilk and M. S. Weiss (2015). "The macromolecular crystallography beamlines at BESSY II of the Helmholtz-Zentrum Berlin: Current status and perspectives." The European Physical Journal Plus **130**(7): 141.

Nanev, C. (2020). "Advancements (and challenges) in the study of protein crystal nucleation and growth; thermodynamic and kinetic explanations and comparison with small-molecule crystallization." Progress in Crystal Growth and Characterization of Materials **66**: 100484.

Neidle, S. (2008). Principles of Nucleic Acid Structure. London, Academic Press.

Nelson, D. L. (2005). Lehninger principles of biochemistry, Fourth edition. New York : W.H. Freeman, 2005.

Nunvar, J., T. Huckova and I. Licha (2010). "Identification and characterization of repetitive extragenic palindromes (REP)-associated tyrosine transposases: implications for REP evolution and dynamics in bacterial genomes." BMC Genomics **11**: 44.

Nuthanakanti, A., I. Ahmed, S. Y. Khatik, K. Saikrishnan and S. G. Srivatsan (2019). "Probing G-quadruplex topologies and recognition concurrently in real time and 3D using a dual-app nucleoside probe." Nucleic Acids Res **47**(12): 6059-6072.

Pan, S., X. Sun and J. K. Lee (2006). "DNA stability in the gas versus solution phases: A systematic study of thirty-one duplexes with varying length, sequence, and charge level." Journal of The American Society for Mass Spectrometry **17**(10): 1383-1395.

Parker, T. M., E. G. Hohenstein, R. M. Parrish, N. V. Hud and C. D. Sherrill (2013). "Quantum-Mechanical Analysis of the Energetic Contributions to π Stacking in Nucleic Acids versus Rise, Twist, and Slide." Journal of the American Chemical Society **135**(4): 1306-1316.

Pettersen, E. F., T. D. Goddard, C. C. Huang, E. C. Meng, G. S. Couch, T. I. Croll, J. H. Morris and T. E. Ferrin (2021). "UCSF ChimeraX: Structure visualization for researchers, educators, and developers." Protein Sci **30**(1): 70-82.

Rentzeperis, D., K. Alessi and L. A. Marky (1993). "Thermodynamics of DNA Hairpins: Contributions of Loop Size to Hairpin Stability and Ethidium Binding." Nucleic Acids Res. **21**(11): 2683-2689.

Rhee, S., Z. Han, K. Liu, H. T. Miles and D. R. Davies (1999). "Structure of a triple helical DNA with a triplex-duplex junction." Biochemistry **38**(51): 16810-16815.

Rhodes, G. (2006). Crystallography Made Crystal Clear: A Guide for Users of Macromolecular Models, Academic Press.

Saenger, W. (1984). Defining Terms for the Nucleic Acids. Principles of Nucleic Acid Structure. New York, NY, Springer New York**:** 9-28.

Saenger, W. (1984). Principles of nucleic acid structure, Springer-Varlag.

Saenger, W., W. N. Hunter and O. Kennard (1986). "DNA conformation is determined by economics in the hydration of phosphate groups." Nature **324**: 385-388.

Schneider, B. and H. M. Berman (1995). "Hydration of the DNA bases is local." Biophys. J. **69**: 2661-2669.

Schneider, B. and H. M. Berman (2006). Basics of Nucleic Acids Structure. Computational Studies of RNA and DNA. J. Sponer and F. Lankas. Dordrecht, Springer**:** 1-44.

Schneider, B., P. Boaeikova, I. Necasova, P. Cech, D. Svozil and J. Cerny (2018). "A DNA structural alphabet provides new insight into DNA flexibility." Acta Crystallogr D Struct Biol **74**(Pt 1): 52-64.

Schneider, B., P. Bozikova, P. Cech, D. Svozil and J. Cerny (2017). "A DNA Structural Alphabet Distinguishes Structural Features of DNA Bound to Regulatory Proteins and in the Nucleosome Core Particle." Genes (Basel) **8**(10).

Schneider, B., K. Patel and H. M. Berman (1998). "Hydration of the DNA Phosphate." Biophysical Journal **75**: 2422-2434.

Seeman, N. C., J. M. Rosenberg and A. Rich (1976). "Sequence specific recognition of double helical nucleic acids by proteins." Proc.Natl.Acad.Sci. USA. **73**: 804-808.

Silverman, S. K. (2004). "Deoxyribozymes: DNA catalysts for bioorganic chemistry." Org Biomol Chem **2**(19): 2701-2706.

Singh, V., B. I. Fedeles and J. M. Essigmann (2015). "Role of tautomerism in RNA biochemistry." RNA **21**(1): 1-13.

Soukup, G. (2003). Nucleic Acids: General Properties.

Spiegel, J., S. Adhikari and S. Balasubramanian (2020). "The Structure and Function of DNA G-Quadruplexes." Trends in Chemistry **2**(2): 123-136.

Vorlickova, M., I. Kejnovska, K. Bednářová, D. Renciuk and J. Kypr (2012). "Circular Dichroism Spectroscopy of DNA: From Duplexes to Quadruplexes." Chirality **24**: 691-698.

Wang, A. H.-J., G. J. Quigley, F. J. Kolpak, J. L. Crawford, J. H. van Boom, G. A. van der Marel and A. Rich (1979). "Molecular structure of a left-handed double helical DNA fragment at atomic resolution." Nature **282**: 680-686.

Wang, J., P. Dong, W. Wu, X. Pan and X. Liang (2018). "High-throughput thermal stability assessment of DNA hairpins based on high resolution melting." J Biomol Struct Dyn **36**(1): 1-13.

Watson, J. D. and F. H. C. Crick (1953). "A structure for deoxyribose nucleic acid." Nature **171**: 737-738.

Weil, J., T. P. Min, C. Yang, S. R. Wang, C. Sutherland, N. Sinha and C. H. Kang (1999). "Stabilization of the i-motif by intramolecular adenine-adenine-thymine base triple in the structure of d(ACCCT)." Acta Cryst. D **55**: 422-429.

Westhof, E. (1988). "Water: an integral part of nucleic acid structure." Ann. Rev. Biophys. Chem. **17**: 125-144.

Zamenhof, S., G. Brawermann and E. Chargaff (1952). "On the Desoxypentose Nucleic Acids from Several Microorganisms." Biochim. Biophys. Acta **9**: 402-405.

Zidek, L., R. Stefl and V. Sklenar (2001). "NMR methodology for the study of nucleic acids." Curr.Opin.Struct.Biol. **11**(3): 275-281.